Research Article

Linh Tran*, Maya Petersen, Joshua Schwab, and Mark J. van der Laan

Robust variance estimation and inference for causal effect estimation

https://doi.org/10.1515/jci-2021-0067
received December 07, 2021; accepted February 14, 2023

Abstract: We present two novel approaches to variance estimation of semi-parametric efficient point estimators of the treatment-specific mean: (i) a robust approach that directly targets the variance of the influence function (IF) as a counterfactual mean outcome and (ii) a modified non-parametric bootstrap-based approach. The performance of these approaches to variance estimation is compared to variance estimation based on the sample variance of the empirical IF in simulations across different levels of positivity violations and treatment effect sizes. In this article, we focus on estimation of the nuisance parameters using correctly specified parametric models for the treatment mechanism in order to highlight the challenges posed by violation of positivity assumptions (distinct from the challenges posed by non-parametric estimation of the nuisance parameters). Results demonstrate that (1) variance estimation based on the empirical IF may provide highly anti-conservative confidence interval coverage (as reported previously), (2) the proposed robust approach to variance estimation in this setting provides conservative coverage, and (3) the proposed modified bootstrap maintains close to nominal coverage and improves power. In the appendix, we (a) generalize the robust approach of estimating variance to marginal structural working models and (b) provide a proof of the consistency of the targeted minimum loss-based estimation bootstrap.

Keywords: estimator variance, influence function, targeted minimum loss-based estimation, asymptotic efficiency, non-parametric bootstrap, positivity assumption, augmented inverse probability-weighted estimation

MSC 2020: 62D20

1 Introduction

A number of estimators are available for the treatment-specific mean outcome parameter (and the corresponding causal contrasts) based on longitudinal data structures, such as inverse probability weighting (IPW) [1,2], double robust augmented IPW (AIPW) [3–8], and targeted minimum loss-based estimation (TMLE) [9]. Variance estimation for each of these estimators is conventionally achieved by using their corresponding influence functions (IFs) [10] based on the empirical distribution or by resampling methods such as the non-parametric bootstrap [11]. However, a number of shortcomings exist with these variance estimation approaches. In particular, the non-parametric bootstrap has been lacking in theory to support its validity and may be computationally prohibitive when machine learning methods are used for the estimation of nuisance parameters. Furthermore, both IF-based and bootstrap-based confidence intervals can become anti-conservative in the setting of positivity violations (i.e., when support for a treatment regime of

* **Corresponding author: Linh Tran,** Department of Statistics, Stanford University, Stanford, California, United States, e-mail: Tranlm@stanford.edu
Maya Petersen, Joshua Schwab, Mark J. van der Laan: Berkeley School of Public Health, University of California, Berkeley, California, United States

interest is minimal for some levels of covariate history [12]). For example, van der Laan and Gruber [9] found IF-based variance estimates for the intervention-specific mean outcome that were anti-conservative when compared with the Monte Carlo variance of the TMLE, leading to poor confidence interval coverage. Petersen et. al. [12] also found poor coverage for IF-based confidence intervals, owing to both practical positivity violations and relatively rare outcomes. Importantly, poor performance of standard variance estimators can occur in finite samples, even when the assumptions for asymptotic validity of these estimators hold [12]. Furthermore, these standard approaches to variance estimation are not sensitive to theoretical violations of the positivity assumptions under which the asymptotic variance of the estimator would be infinity, i.e., when the positivity assumption needed for identification fails due to lack of support in the underlying data-generating process. Improved approaches to variance estimation in the context of positivity violations are thus needed that are able to (1) serve as a "red flag" to alert the analyst if the data at hand provide insufficient information to estimate the desired causal parameter with any reasonable degree of accuracy, and (2) provide closer to nominal confidence interval coverage and type 1 error control in these challenging estimation settings.

Previous studies [12,13] proposed estimating the asymptotic variance of the estimator with a parametric bootstrap based on a fit of the density of the data-generating distribution, involving estimation of individual factors of the likelihood. This proposal corresponds with evaluation of the variance of a given estimator using the data at hand as a given data-generating experiment. The consistency of this estimator relies on a consistent estimation of the corresponding factors of the likelihood. This parametric bootstrap integrates over sparse events and therefore will explode the variance. An extremely large sample is therefore needed to get the true variance under this Monte Carlo scheme. As a consequence, this parametric bootstrap-based variance estimate was only proposed as a measure to raise a red flag for unreliable statistical inference due to poor data support. In addition, in the context of sparsity in order to obtain a valid estimate of the variance, one needs to (i) sample a large number of bootstrap samples and (ii) refit the likelihood in each iteration in order to capture the rare observations that, nonetheless, heavily contribute to the variance. Thus, the bootstrap method is extremely computer-intensive, making this Monte Carlo scheme an intractable method for complex estimators, particularly those that rely on machine learning for nuisance parameter estimation.

In this article, we present two approaches to variance estimation in the context of positivity violations to address these challenges. We use analytic expressions to compute the variance of the efficient influence function (EIF) for the statistical target parameter corresponding (under assumptions) to the counterfactual mean outcome under a longitudinal dynamic treatment regime, thereby providing the asymptotic variance of estimators solving the estimating equation corresponding to this function. These analytic expressions naturally integrate over the rare observations and thereby avoid the finite sample bias in variance estimation using standard influence curve or non-parametric bootstrap-based methods due to the rare aforementioned observations. With this, we construct robust plug-in-type estimators of these asymptotic variances that are consistent if both the treatment mechanism and treatment-specific means of specified outcomes are consistently estimated (as our derived expression requires both). These estimators require estimation of the treatment mechanism and several treatment-specific means of specified outcomes (defined as a function of the observed data structure, indexed by the estimator of the treatment mechanism), which can be estimated with either an estimating equation-type IPW estimator or an efficient substitution-based method such as a TMLE [9]. The resulting variance estimator, unlike current alternatives based on taking the variance of the empirical IF, or using a non-parametric bootstrap, will become very large whenever the estimated treatment mechanism reflects practical or theoretical violations of the positivity assumption.

While this newly presented approach performs well in estimating the asymptotic variance of estimators solving the estimating equation corresponding to the EIF, a lower finite sample variance should be expected for substitution-based estimators such as TMLE [9], due to the guaranteed parameter boundaries provided by the estimator. We therefore additionally present a second bootstrap-based approach of estimating the finite sample variance. This bootstrap improves on the estimator based on the empirical variance of the variance of the IF by making use of rare observations in the update step of the estimator. Importantly, it does not require re-estimation of the individual factors of the likelihood and therefore reduces the computational burden compared to both the standard parametric and non-parametric bootstrap methods. The

resulting reduction in the computational load (compared to a fully non-parametric bootstrap approach, which refits the likelihood for each iteration) allows for a more tractable approach at estimating the variance. Furthermore, the modified bootstrap is asymptotically consistent under reasonable assumptions, namely, the same essential assumptions needed for the estimator of the target parameter itself to be asymptotically linear. In other words, the bootstrap we propose is valid whenever the original TMLE is asymptotically linear. A non-parametric full bootstrap can easily break down due to the machine learning algorithms behaving differently under sampling from P_n than under P_0, where P_n is our empirical distribution and P_0 is the true distribution. Thus, while the non-parametric bootstrap is not consistent for these efficient estimators using machine learning, our proposed bootstrap is consistent.

1.1 Organization of this article

In Section 2, we formally define the observable data, likelihood, and statistical model for its distribution. Our target parameter of the treatment-specific mean outcome is defined, along with its EIF. We briefly review the causal model and identification assumptions under which this statistical parameter of the observed data distribution corresponds with the desired causal parameter of the counterfactual distribution, along with the currently common approach of IF-based estimator variance estimation.

Section 3 presents an approach for robust estimation of the variance of the EIF under sparsity. The expression for the variance of the EIF is presented along with both IPW and TMLE-based approaches for estimating this parameter. To illustrate, an example is given for a point treatment setting under a static treatment regime. Advantages of this approach, implemented in the ltmle R package [12,14] but not previously described in the literature, are covered. Appendix A generalizes the approach to working marginal structural working models and provides proofs.

Section 4 discusses a second approach of estimating the estimator variance using the bootstrap under a modified TMLE. Appendix B proves the consistency of this bootstrap estimator.

Section 5 illustrates the performance of the variance estimators presented in Sections 3 and 4 by applying them in simulations to both a single time point and longitudinal treatment settings to estimate the variance of an (iterated conditional expectation) TMLE point estimator [9,15] under varying effect sizes and degrees of positivity violation. We focus on the setting in which the treatment mechanism is estimated according to a correctly specified parametric model in order to distinguish challenges to inference due to positivity from the distinct challenges to variance estimation and inference due to potential slow convergence of machine learning-based nuisance parameter estimators. We discuss extensions to the setting of machine learning-based nuisance parameter estimation in the discussion. Results show that, in settings of substantial positivity violation, the standard empirical IF-based approach results in anti-conservative confidence interval coverage. In contrast, the robust approach provides conservative coverage, and thus an effective diagnostic for settings in which standard approaches may result in misleading inference. Finally, the proposed bootstrap approach provides the closest to nominal coverage of the three estimators, and maintains higher power than the robust approach.

We conclude with a discussion in Section 6, which reviews the results, benefits of this new approach, potential limitations, and future directions.

2 The causal roadmap: the statistical estimation problem and causal identification

Consider a longitudinal study in which subjects are seen at each time point t from $t = 0, 1, \ldots, K + 1$. The observable data structure on a randomly sampled subject is

$$O = (L(0), A(0), L(1), A(1), \ldots, A(K), \quad Y = L(K+1)) \overset{iid}{\sim} P_0, \tag{1}$$

where $L(0)$ includes all baseline covariates, $A(t)$ denotes an intervention (or treatment) node at time t, and $L(t)$ denotes all time-varying covariates at time point t, measured between the intervention nodes $A(t^-)$ and $A(t)$, where for notational convenience, we define $t^- \equiv t-1$. Our outcome of interest $Y = L(K+1)$ is an outcome measured after the final intervention $A(K)$. We observe n independent and identically distributed (iid) copies $O_i : i = 1, \ldots, n$, of O.

The likelihood $L(O)$ for the observable data is the product of conditional probabilities such that the likelihood for subject i is

$$\begin{aligned} L(O_i) &= p_0(L_i(0), A_i(0), L_i(1), A_i(1), \ldots, L_i(K+1)) \\ &= p_0(L_i(K+1)|\bar{L}_i(K), \bar{A}_i(K)) \cdot p_0(A_i(K)|\bar{L}_i(K), \bar{A}_i(K-1)) \\ &\quad \cdot p_0(L_i(K)|\bar{L}_i(K-1), \bar{A}_i(K-1)) \cdot p_0(A_i(K-1)|\bar{L}_i(K-1), \bar{A}_i(K-2)) \cdots p_0(L_i(0)) \\ &= \left[\prod_{t=0}^{K+1} \underbrace{p_0(L_i(t)|\bar{L}_i(t^-), \bar{A}_i(t^-))}_{q_{0,t}(L(t))|Pa(L(t))}\right] \cdot \left[\prod_{t=0}^{K} \underbrace{p_0(A_i(t)|\bar{L}_i(t), \bar{A}_i(t^-))}_{g_{0,t}(A(t))|Pa(A(t))}\right], \end{aligned} \tag{2}$$

where $\bar{X}(t) \equiv (X(1), X(2), \ldots, X(t))$, $A(-1) = L(-1) = \varnothing$, and $p_0(o)$ denotes $p_0(O = o)$ under the true distribution P_0 where we assume O is discrete for the sake of presentation.

The statistical model $\mathcal{M}$ for the data involves assumptions, if any, only on the conditional distributions of $A(t)$, given $Pa(A(t)) = (\bar{L}(t), \bar{A}(t^-))$, $t = 0, \ldots, K$. Let

$$P_0^d(l) \equiv \prod_{t=0}^{K+1} P_{0,L(t)}(l(t))|\bar{l}(t^-), d(\bar{l}(t^-)),$$

denote the G-computation formula for the post-intervention distribution of an intervention that sets $\bar{A}(K) = d(\bar{l}(K))$ [16]. We use the notation $P_{L(t)}$ for a conditional distribution of $L(t)$, given $Pa(L(t)) = (\bar{L}(t^-), \bar{A}(t^-))$. Let $L^d = (L(0), \ldots, Y^d = L^d(K+1))$ be a random variable under the post-intervention distribution P_0^d. The statistical target estimand is defined here as $\Psi(P_0) = \mathbb{E}_{P_0^d}[Y^d]$, i.e., the mean of the outcome at time $K+1$ under this distribution. We note that $\Psi : \mathcal{M} \to \mathbb{R}$ represents a target parameter mapping on the statistical model to the real line. Defining $t^+ \equiv t+1$, the EIF of Ψ at P is given by previous studies [6,15,17]

$$D^*(P)(O) = \sum_{t=0}^{K+1} D_t^*(P)(O),$$

where

$$\begin{aligned} D_0^*(P)(L(0)) &= \bar{Q}_1^d - \bar{Q}_0^d \\ D_t^*(P)(\bar{A}(t^-), \bar{L}(t^-)) &= H_t(g)\left(\bar{Q}_{t^+}^d - \bar{Q}_t^d\right) : t = 1, 2, \ldots, K+1, \end{aligned}$$

where

$$\begin{aligned} H_t(g) &= \frac{\mathbb{I}(\bar{A}(t^-) = d(\bar{l}(t^-)))}{g_{0:t^-}(\bar{A}(t^-), \bar{L}(t^-))}, \\ \bar{Q}_{K+2}^d &= Y, \\ \bar{Q}_t^d &= \mathbb{E}_P[Y^d|\bar{L}^d(t^-) = \bar{L}(t^-)] : t = 1, 2, \ldots, K+1, \\ \bar{Q}_0^d &= \mathbb{E}_P[Y^d]. \end{aligned} \tag{3}$$

Here, $g_{0:t^-}(\bar{A}(t^-), \bar{L}(t^-))$ denotes the cumulative probability of treatment up to time $t-1$, i.e.,

$$g_{0:t^-}(\bar{A}(t^-), \bar{L}(t^-)) \equiv \prod_{u=0}^{t-1} p(A(u)|\bar{L}(u), \bar{A}(u-1)).$$

Furthermore, $\bar{Q}_t^d$ is defined by recursive regression, starting at $t = K + 1$ and moving backward in time. For notational convenience, we let $H_0 = 1$ so that

$$D^*(P)(O) = \sum_{t=0}^{K+1} H_t(g)\left(\bar{Q}_{t^+}^d - \bar{Q}_t^d\right).$$

2.1 Causal model

Under additional assumptions about the data-generating process, the target statistical estimand $\Psi(P_0)$ is equal to the mean of the counterfactual outcome Y_d under an intervention to set the vector of treatment nodes to value $d(\bar{l}(K))$ (i.e., the counterfactual outcome under a specified dynamic regime). Specifically, for interventions of interest $d \in \mathcal{D}$, we assume sequential randomization [16]

$$Y_d \perp\!\!\!\perp A(t) | \bar{L}(t), \bar{A}(t^-) : t = 0, 1, \ldots, K$$

and positivity [18]

$$P(\bar{A}(t) = d(\bar{L}(t)) | \bar{L}(t), \bar{A}(t^-) = d(\bar{L}(t))) > 0 \quad \text{a.e.} \quad : t = 0, 1, \ldots, K. \tag{4}$$

Regarding the assumption of positivity, we note that as $P(\bar{A}(t) = d(\bar{L}(t)) | \bar{L}(t), \bar{A}(t^-) = d(\bar{L}(t))) \to 0$, we have that $H_t(g) \to \infty$ resulting in $\text{var}[D^*(P)(O)] \to \infty$; following the literature, we refer to violation of the positivity assumption (4) as a theoretical positivity violation and near violations resulting in lack of adequate support in finite samples as practical positivity violations.

2.2 Review of IF-based variance

Recall that an estimator $\hat{\Psi}(P_n)$ is considered to be asymptotically linear if and only if

$$\hat{\Psi}(P_n) - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} D(P_0)(O_i) + o_p(n^{-1/2})$$

for some mean 0 finite variance IF $D(P_0)(O)$ [10]. If an estimator is asymptotically linear, then it will be asymptotically normal with variance equal to the variance of the IF over n. The asymptotic variance of the estimator can therefore be consistently estimated with the variance of the empirical IF $D(P_n)(O)$, i.e., $\hat{\text{var}}[\hat{\Psi}(P_n)] = \text{var}[D(P_n)(O)]/n$, which implies an asymptotically valid confidence interval (assuming the theoretical positivity assumption (4)). However, as we discuss in the following sections, this approach has substantial limitations both as a diagnostic that the theoretical positivity assumption fails, and for finite sample inference in the context of practical positivity violations.

2.2.1 TMLE

One possible estimator that solves the estimating equation corresponding to the EIF for intervention-specific mean outcomes is TMLE [19]. Naive plug-in estimators can be too biased with a dominating term being the empirical mean of the EIF, and solving the estimating equation can reduce this bias.

This plug-in estimator works by forming an initial fit of the q_0 factors of the likelihood and subsequently perturbing it such that the estimating equation corresponding to the EIF is solved. We assume the use of the iterated conditional outcome TMLE [9,15] but modified to improve robustness to data sparsity by incorporation of the inverse probability of treatment as a weight rather than a covariate in the targeted update [20–22]. We restrict ourselves to this estimator for our estimation problem, such that our attention is focused on estimation of the estimator's variance. Note that our proposed variance estimators also apply to estimating equation approaches, such as the double robust AIPW [3–8].

3 Semi-targeted estimation of the EIF variance

A common approach to variance estimation and corresponding inference based on this point estimator is thus based on taking the sample variance of the empirical EIF divided by sample size, using either the initial or the targeted empirical estimates of the q_0 factors of the likelihood.

The basis of the proposed "robust" approach to variance estimation is to express the variance of $D(P_0)(O)$ as an expectation, allowing us to estimate the variance as a mean and the target estimation of the variance directly using TMLE. The following describes how to obtain a TMLE of the variance of each component of the EIF σ_t^2 in the setting of a scalar parameter. We provide a proof for the more general working MSM setting in Appendix A for the interested reader. Both robust estimators are implemented as options in the ltmle R package.

3.1 Expression for variance of the EIF for $\mathbb{E}Y_d$

Under regimens $d(\bar{l}(K))$, we have

$$\sigma_0^2 \equiv \mathbb{E}_0[D^*(P_0)(O)]^2 = \sum_{t=0}^{K+1} \mathbb{E}_0\left[H_t^2(g_0)\left(\bar{Q}_{0,t^+}^d - \bar{Q}_{0,t}^d\right)^2\right],$$

where g_0 represents the true cumulative probability of treatment up to time $t-1$, i.e., $g_{0:t^-}$ under P_0.

Using the expression for $H_t(g)$ from equation (3), and first taking the conditional expectation w.r.t. $\bar{A}(t^-)$ given $X = (\bar{L}^d : d)$, it follows that this can be written as:

$$\sigma_0^2 = \sum_{t=0}^{K+1} \mathbb{E}_{P_0^d}\left[\frac{(\bar{Q}_{0,t^+}^d - \bar{Q}_{0,t}^d)^2(\bar{L}^d(t))}{g_0(d(\bar{l}(t^-)), \bar{L}^d(t^-))}\right], \tag{5}$$

where we define $g_0(d(\bar{l}(-1)), \bar{L}^d(-1)) = 1$ so that the term at $t = 0$ equals $\mathbb{E}_{L(0)}[\bar{Q}_{0,1}^d(L(0)) - \mathbb{E}_0 Y^d]^2$. This is simply a sum of expectations over $t \in \{0, 1, \ldots, K+1\}$. For notational convenience, we re-write equation (5) as

$$\sigma_0^2 = \sum_{t=0}^{K+1} \sigma_t^{2,d} = \sum_{t=0}^{K+1} \mathbb{E}_{P_0^d}[S_t^d(\bar{Q}_0, g_0)(\bar{L}^d(t))], \tag{6}$$

for the specified function

$$S_t^d(\bar{Q}_0, g_0)(\bar{L}^d(t)) \equiv \frac{(\bar{Q}_{0,t^+}^d - \bar{Q}_{0,t}^d)^2(\bar{L}^d(t))}{g_0(d(\bar{l}(t^-)), \bar{L}^d(t^-))} : t = 0, 1, \ldots, K+1.$$

Note that given $(\bar{Q}_0, g_0)$, we have that $\mathbb{E}_{P_0^d}[S_t^d(\bar{Q}_0, g_0)]$ is the mean of a counterfactual $S_t^d(\bar{Q}_0, g_0)(\bar{L}^d(t))$, i.e., the mean of a real-valued function (indexed by $d(\bar{l})$ itself) of $\bar{L}^d(t)$, which needs to be estimated based on the longitudinal data structure $L(0), A(0), \ldots, A(t-1), L(t)$. Given $\bar{Q}_0, g_0$, we observe the outcome $S_t^d(\bar{Q}_0, g_0)(\bar{L}_i(t))$, $i = 1, 2, \ldots, n$, so that we can represent the observed data structure as $L(0), A(0), \ldots, A(t-1), S_t^d(\bar{Q}_0, g_0)(\bar{L}(t))$, and we wish to estimate the statistical target parameter

$$\mathbb{E}_{P_0^d}[S_t^d(\bar{Q}_0, g_0)] = \sum_{\bar{l}(t)} S_t^d(\bar{Q}_0, g_0)(\bar{l}(t)) P_0^d(\bar{L}^d(t) = \bar{l}(t)) : t = 0, 1, \ldots, K+1, \tag{7}$$

where again we assume $l(t)$ is discrete for the sake of presentation.

3.1.1 Estimation of variance of the EIF

With the expression for the variance of the EIF in hand (equation (6)), we can now form estimators that target this parameter. $\bar{Q}_0$ and g_0 are not known in practice, though estimates $\bar{Q}_n^*$ and g_n will be readily available if estimating $\mathbb{E}[Y_d]$ using a double robust estimator such as TMLE, thus providing us with the

observed outcome $S_t^d(\bar{Q}_n^*, g_n)(\bar{L}(t))$. Treating this variable as our new time point-specific outcome, our goal is to estimate the mean of this variable over the post-intervention distribution of $\bar{L}^d(t)$. For notational convenience, let $Z^d(t) \equiv S_t^d(\bar{Q}_0, g_0)(\bar{L}(t))$ represent the observable outcome and $(L(0), A(0), \ldots, A(t-1), Z^d(t))$ represent the observed data structure.

One possible approach to estimating each of the components (equation (7)) is to use a simple IPW estimator [1]

$$\hat{\sigma}^{2,d}_{t,n,\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(\bar{A}_i(t^-) = d(\bar{l}(t^-)))}{g_{0:t^-,n}(d(\bar{l}(t^-)), \bar{L}_i(t^-))} Z_n^d(t),$$

where $Z_n^d(t) = S_t^d(\bar{Q}_n, g_n)(\bar{L}(t))$. However, such an estimator would still be subject to underestimation of the variance by ignoring the contribution of observations that selected a likely treatment $\bar{A}_i$, even though their probability of following $d(\bar{l})$ is very small. In other words, subjects i with small probabilities of following $d(\bar{l})$ would be unlikely to be observed with $\bar{A}_i = d(\bar{l})$ resulting in an indicator value of 0 for the numerator and, consequently, a contribution of 0 to the IPW estimator. Therefore, we stress that it is important to use a plug-in estimator such as TMLE [9] to estimate this parameter. A plug-in estimator will integrate over all $\bar{l}(t)$ in the support of $P_{t,n}^d$ and thus contribute many large values of $S_{t,n}^d(\bar{Q}_n^*, g_n)$ when there are practical or theoretical positivity assumption violations. In addition, the TMLE is a double robust estimator (for the point estimate) so that it will yield a consistent estimator of this variance if g_n is consistent for the true g_0 or $\bar{Q}_n$ is consistent for $\bar{Q}_0$.

Given $\bar{Q}_0$ and g_0, we will now provide a succinct summary of the TMLE of $\sigma^{2,d}_{0,t} = \mathbb{E}_{P_0^d}[Z^d(t)]$ that is based on the iterative sequential regression approach. Note that this iterative sequential regression approach is similar to the one presented by van der Laan and Gruber [9] for the intervention-specific mean outcome parameter. Denote the counterfactual of $Z^d(t)$ under treatment d' with $Z^{d,d'}(t)$. Recall that we observe $S_t^d(\bar{Q}_0, g_0)(\bar{L}(t))$, whereas we wish to estimate $S_t^d(\bar{Q}_0, g_0)(\bar{L}^{d'}(t))$ (n.b. here, we focus only on when $d' = d$). Let $P_0^{d'}$ be the G-computation formula [16] corresponding with this intervention $\bar{A}(t^-) = d'(\bar{l}(t^-))$. We wish to estimate $\sigma^{2,d}_{0,t} = \mathbb{E}_{P_0^d}[Z^{d,d'}(t)]$, which can be represented as a series of iterated conditional expectations

$$\sigma^{2,d}_{0,t} = \mathbb{E}[\mathbb{E}[\cdots \mathbb{E}[\mathbb{E}[Z^d(t)|\bar{L}^d(t-1)]|\bar{L}^d(t-2)]\cdots|\bar{L}^d(0)]].$$

The EIF for this target parameter $\sigma_t^{2,d}$ is given by:

$$D^*_{\sigma_t^{2,d}}(P)(O) = \sum_{m=0}^{t} H_m^{d,t}(g)\left(\bar{Q}_{m^+}^{d,\sigma_t^2} - \bar{Q}_m^{d,\sigma_t^2}\right),$$

where we define

$$\begin{aligned}
\bar{Q}_{t+1}^{d,\sigma_t^2} &= Z^d(t) \\
\bar{Q}_m^{d,\sigma_t^2} &= \mathbb{E}_P[Z^d(t)|\bar{L}^d(m) = \bar{L}(m)] : m = 1, 2, \ldots, t \\
H_m^{d,t}(g) &= \frac{\mathbb{I}(\bar{A}(m^-) = d(\bar{l}(m^-)))}{g_{0:m^-}(d(\bar{l}(m^-))\bar{L}(m^-))} : m = 1, 2, \ldots, t \\
H_0^{d,t} &= 1.
\end{aligned}$$

Therefore, the EIF for $\sigma^2 = \sum_t \sigma_t^{2,d}$ is simply $D^*_{\sigma^2} = \sum_t D^*_{\sigma_t^{2,d}}$.

With the EIF established, the TMLE of $\sigma_t^{2,d}$ (in similar fashion to van der Laan and Gruber [9]) is now defined as follows.

- Estimates $g_{0:m^-,n} : m = 1, 2, \ldots, t$ are readily available if estimating $\mathbb{E}[Y_d]$ using an estimator, which solves the estimating equation corresponding to the EIF such as TMLE.
- Set $\bar{Q}_{t,n}^{d,\sigma_t^2} = Z_i^d(t)$. Determine the range (a, b) for $Z_i^d(t)$, $i = 1, \ldots, n$ and target this initial fit using a parametric submodel respecting this range (a, b) by adding the clever covariate $H_t^{d,t}$ (on, say, the logistic scale), using the initial fit $\bar{Q}_{t,n}^{d,\sigma_t^2}$ as off-set. The resulting updated fit is denoted with $\bar{Q}_{t,n}^{d,\sigma_t^2,*}$.

- Given $\bar{Q}_{t,n}^{d,\sigma_t^2,*}$, we can recursively for $m = t-1, t-2, \ldots, 1$:
  - Regress the targeted fit $\bar{Q}_{m^+,n}^{d,\sigma_t^2,*}$ onto $\bar{A}(m^-) = d(\bar{l}(m^-)), \bar{L}(m^-)$, using logistic regression to respect the range (a, b). Denote the fit $\bar{Q}_{m,n}^{d,\sigma_t^2}$.
  - Target this initial fit respecting the range (a, b) with clever covariate $\mathbb{I}(\bar{A}(m^-) = d(\bar{l}(m^-)))$ and observational weight $\frac{1}{g_{0:m^-}(d(\bar{l}(m^-)),\bar{L}(m^-))}$ (on the logistic scale), and denote this targeted fit of $\bar{Q}_m^{d,\sigma_t^2}$ with $\bar{Q}_{m,n}^{d,\sigma_t^2,*}$. Note that we already have $g_{0:m^-}(d(\bar{l}(m^-)), \bar{L}(m^-))$ readily available from Step 1. Furthermore, we have the ability to use different submodels (e.g., using a linear submodel that incorporates the propensity scores into the clever covariate, rather than as weights). We defer discussing this until Section 5.2.1.
- At $m = 1$, we have the estimate $\bar{Q}_{1,n}^{d,\sigma_t^2,*}$, which now is a function of only $L(0)$. Finally, we take the average of $\bar{Q}_{1,n}^{d,\sigma_t^2,*}$ w.r.t. the empirical distribution of $L_i(0)$, i.e., $\bar{Q}_{0,n}^{d,\sigma_t^2,*} = \frac{1}{n}\sum_{i=1}^n \bar{Q}_{1,n}^{d,\sigma_t^2,*}(L_i(0))$. The resulting $\hat{\sigma}_{t,n,\text{TMLE}}^{2,d} = \bar{Q}_{0,n}^{d,\sigma_t^2,*}$ is the desired TMLE of $\sigma_t^{2,d}$.

We refer any readers unfamiliar with TMLE to Gruber and van der Laan [23] for further details.

3.1.2 Application to single time point treatment setting

For the sake of illustration, let us consider the method presented earlier for the estimation of the variance of the EIF for the case that $O = (L(0), A(0), Y = L(1))$ and the target parameter is $\mathbb{E}[Y^a]$ for a static point treatment a.

In this case, the variance of the EIF is represented as:

$$\begin{aligned}
\sigma_0^2 &= \mathbb{E}_0[D^*(P_0)(O)]^2 \\
&= \mathbb{E}_0\left[\frac{\mathbb{I}(A=a)}{g_0(a|L(0))}(Y - \bar{Q}_0^a(L(0))) + \bar{Q}_0^a(L(0)) - \mathbb{E}[Y^a]\right]^2 \\
&= \mathbb{E}_0\left[\frac{\mathbb{I}(A=a)}{g_0(a|L(0))}(Y - \bar{Q}_0^a(L(0)))\right]^2 + \mathbb{E}_0[\bar{Q}_0^a(L(0)) - \mathbb{E}[Y^a]]^2 \\
&= \mathbb{E}_{P_0^a}\left[\frac{(Y^a - \bar{Q}_0^a(L(0)))^2}{g_0(a|L(0))}\right] + \mathbb{E}_0[\bar{Q}_0^a(L(0)) - \mathbb{E}[Y^a]]^2.
\end{aligned} \tag{8}$$

If using an estimator that solves the estimating equation corresponding to the EIF for the estimation of $\mathbb{E}[Y^a]$ such as TMLE, we are provided with estimators g_n and $\bar{Q}_n^*$ of $g_0(A|L(0))$ and $\bar{Q}_0^a(L(0)) = \mathbb{E}[Y^a|L(0)] = \mathbb{E}_0[Y|A=a, L(0)]$, respectively. The second term in the final expression of equation (8) is easily estimated with the empirical distribution. Given g_0 and $\bar{Q}_0$, the first term can be represented as the mean of a counterfactual $S^a(L^a(0)) \equiv (Y^a - \bar{Q}_0^a(L(0)))^2/g_0(a|L(0))$, which needs to be estimated based on $(L(0), A, S^a(L(0), Y))$, where $S^a(L(0), Y) = (Y - \bar{Q}_0(a, L(0)))^2/g_0(a|L(0))$ represents the observed outcome. For example, we can use a TMLE estimator $\mathbb{E}_n^*[S^a(L(0), Y^a)]$ of $\mathbb{E}_0[S^a(L(0), Y^a)] = \mathbb{E}_{L(0),0}[\mathbb{E}_0[S^a|A=a, L(0)]]$. The TMLE estimate $\mathbb{E}_n^*[S^a|A=a, L(0)]$ of $\mathbb{E}_0[S^a|A=a, L(0)]$ is defined by determining the range (a, b) of $S^a(L_i(0), Y_i)$, obtaining an initial regression fit of $\mathbb{E}_0[S^a|L(0), A]$ that respects this range, representing it as a logistic regression fit bounded by (a, b), and updating the latter by fitting a univariate logistic regression with clever covariate $\mathbb{I}(A=a)$ and observational weight $1/g_0(a|L(0))$, using the initial fit as an off-set. Regarding the initial fit $\mathbb{E}_n[S^a|A=a, L(0)]$, recall from the aforementioned fact that S^a is a function of $L(0)$, which results in the initial fit being exactly $(Y - \bar{Q}_0(a, L(0)))^2/g_0(a|L(0))$. We can therefore use $(Y - \bar{Q}_n(a, L(0)))^2/g_n(a|L(0))$, thus not requiring that we perform additional regression. Following the update step, the TMLE of $\mathbb{E}_0[S^a(L(0), Y^a)]$ is now given by $\frac{1}{n}\sum_{i=1}^n \mathbb{E}_n^*[S^a|L_i(0), A=a] = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Q}_n^*(a, L_i(0)))^2/g_n(a|L_i(0))$, so that

$$\hat{\sigma}_n^{2,*} = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \bar{Q}_n^*(a, L_i(0)))^2}{g_n(a|L_i(0))} + \frac{1}{n}\sum_{i=1}^{n}(\bar{Q}_n^*(L_i(0), a) - \hat{\psi}_n^*)^2,$$

where $\hat{\psi}_n^*$ is the targeted estimate of $\mathbb{E}[Y^a]$.

3.2 Advantages of this plug-in estimator of the asymptotic variance of the EIF

Since σ_0^2/n equals the asymptotic variance of an asymptotically efficient estimator, it provides a good measure of the amount of information in the data for the target parameter of interest. Therefore, it is sensible to view σ_0^2/n as a measure of sparsity for the target parameter of interest. If g_n is a good estimator of g_0, then our proposed plug-in estimator $\hat{\sigma}_n^2$ is much less subject to underestimation due to sparsity than currently available estimators such as the sample variance of the estimated IF, and the bootstrap-based estimate of the variance of an efficient estimator. This plug-in estimate $\hat{\sigma}_n^2$ represents a variance of the estimate of the EIF, which involves the integration of rare combinations of treatment and covariates that are unlikely to occur in the actual sample.

In particular, if there are theoretical violations of the positivity assumption, then this true variance σ_0^2 equals infinity, and if g_n approximates g_0 well, then also the estimate $\hat{\sigma}_n^2$ will generate very large values, demonstrating the lack of identifiability and thereby raising a red flag for finite sample sparsity bias in the estimators (beyond the large confidence intervals generated by $\hat{\sigma}_n^2$).

We note that a disadvantage is that it is overly sensitivity to positivity violations by getting too large relative to real finite sample variance of the TMLE.

4 Variance estimation for substitution-based estimators

The plug-in estimator of the asymptotic variance of the EIF presented earlier is superior to the more common approach of taking the empirical EIF variance over the sample (i.e., $\text{var}[D^*(P_n)(O)]$), in that there is a much stronger contribution of combinations of treatment and covariates that are unlikely to occur in the actual sample. In finite samples, however substitution-based estimators such as TMLE (which are guaranteed to solve the EIF within a bounded range) are often observed to have smaller variance than their asymptotic variance. This is due to the mere fact that they are guaranteed to respect the global constraints of the statistical model and target parameter mapping. That is, as opposed to estimators defined as the solution to estimating equations, which tend to result in estimates outside parameter boundaries as the EIF variance increases, the use of substitution estimators in finite samples will often retain an estimator variance that is smaller than the EIF variance divided by the sample size, n. Thus, using the newly presented robust EIF variance method can result in overestimation of the estimator variance for these types of estimators. This therefore motivates us to develop an estimator of variance that is less conservative, i.e., more aligned to the true variance of substitution-based estimators (such as TMLE), particularly in settings of positivity violations.

One alternative approach for estimating the variance for substitution-based estimators is to conduct a non-parametric bootstrap. The n observations are sampled with replacement and used to form an estimate of the parameter over B iterations. However, as stated earlier, the non-parametric bootstrap is generally invalid and not theoretically supported. Additionally, this is a very computer-intensive method that usually requires estimating the full likelihood (i.e., P_0^d) of the longitudinal data structure within each sampled iteration and is therefore normally infeasible in practice unless conducted within an *a priori* selected smaller parametric statistical model such as logistic regression.

In this section, we present an alternative bootstrap-based approach that, unlike the standard non-parametric bootstrap, is both computationally feasible and theoretically valid. That is, this bootstrap approach allows us to estimate the variance of the estimator while avoiding re-estimation of g_0 and $\bar{Q}_0$.

To facilitate this, we propose a modification of the usual TMLE such that the targeting step is separated from the initial estimation of $\bar{Q}_0$. Recall that the typical TMLE, as implemented, pivots between the targeting step and the initial estimator for the next regression (preventing us from separating the initial fit from the targeting step). We propose a minor modification of the TMLE that separates these steps, first estimating all of the initial regressions and subsequently targeting the fits in a separate step. This modified TMLE can then be bootstrapped via only the targeting step. We provide a proof of its consistency in Appendix B. Note that, because the modified TMLE has the same asymptotic behavior as the original TMLE, the bootstrap is theoretically supported and will continue to have valid inference. To ensure that this does not result in anti-conservative behavior, we use a bootstrap with a TMLE update that is non-robust (i.e., by defining the clever covariate $H_t(g)$ with the denominator g_n), thereby resulting in large values for observations that are highly unlikely to follow the treatment regime of interest given their covariate history, even when they in fact fail to do so in the sample.

4.1 Modified TMLE for $\mathbb{E}[Y^d]$

To reduce the computational burden that bootstrapping requires, we first present the modified TMLE approach for generating a point estimate of the parameter $\mathbb{E}[Y^d]$. This parameter can be estimated by the following steps:

(1) Estimate $g_{0:t^-}(\bar{A}, \bar{L}) : t = 1, 2, \ldots, K+1$ and denote the fits $g_{0:t^-,n}$.

(2) Determine the range (a, b) for $\mathbb{E}[Y^d]$. Recursively for $t = K+1, K, \ldots, 1$, estimate the conditional expectation $\bar{Q}_t^d = \mathbb{E}[\bar{Q}_{t^+}^d | \bar{L}(t^-), \bar{A}(t^-) = d(\bar{l}(t^-))]$ respecting this range. Denote the fits $\bar{Q}_{t,n}^d$. *We stress that this step is crucially different than the typical TMLE, in that all of the initial regression fits are done simultaneously*.

(3) For time $t = K+1$, target the initial fit $\bar{Q}_{K+1,n}^d$ by using a parametric submodel respecting the range (a, b) by adding the covariate $\mathbb{I}(\bar{A}(K) = d(\bar{l}(K)))$ and observational weight $1/g_{0:K,n}$ (on the logistic scale), using the initial fits as off-set, and setting Y as the dependent variable. Denote this updated fit as $\bar{Q}_{K+1,n}^{d,*}$.

(4) Given $\bar{Q}_{K+1,n}^{d,*}$, we can recursively for $t = K, K-1, \ldots, 1$ target the initial fits $\bar{Q}_{t,n}^d$ by using parametric submodels respecting the range (a, b), adding the covariates $\mathbb{I}(\bar{A}(t^-) = d(\bar{l}(t^-)))$ and observational weight $1/g_{0:t^-,n}$ (on the logistic scale), using the initial fits as off-set, and setting $\bar{Q}_{t^+,n}^{d,*}$ as the dependent variable. Denote the updated fits as $\bar{Q}_{t,n}^{d,*}$.

(5) At $t = 1$, we have the estimate $\bar{Q}_{1,n}^{d,*}$, which now is a function of only $L(0)$. Taking the average of $\bar{Q}_{1,n}^{d,*}$ w.r.t. the empirical distribution of $L_i(0)$ gives us the desired TMLE estimate of $\mathbb{E}[Y^d]$.

This estimator also solves the EIF and is therefore also asymptotically linear and efficient. We note that the analysis of this TMLE is identical to the typical TMLE presented by van der Laan and Gruber [9], with the only difference being the initial estimator fits. Here, the initial estimators are the original ones, whereas the previous TMLE is implemented with initial estimators using the targeted fits for the outcome.

We emphasize that this estimator is proposed for the sake of the bootstrap method for variance estimation. It is recursive, in that each fit, $\bar{Q}_{t,n}^d$ is dependent upon the fit at t^+. As opposed to the TMLE, the recursive nature of this TMLE is self-contained within each step. In other words, each estimation step in this TMLE can be performed independently of the other steps. This allows the analyst to form all of the initial fits P_n prior to performing any of the targeted updates.

4.2 Bootstrapping the modified TMLE

The new TMLE approach presented above can be bootstrapped in a fully non-parametric manner, such that observations are drawn with replacement prior to fitting the full-likelihood P_0^d and used to form an estimate

of the parameter, leading to an estimate of estimator variance. Our recommendation is to only bootstrap the targeting step. More specifically, once the fits $g_{0:t^-,n}$ and $\bar{Q}^d_{t,n}$ are formed for $t = 1, 2, \ldots, K+1$, steps 3–5 are carried out in the bootstrap such that for $b = 1, 2, \ldots, B$, we have

$$Q^*_{n,b} = Q_n(\varepsilon_b)$$

for a user-selected submodel $P(\varepsilon)$. The estimator variance is then estimated by taking the variance over the bootstrapped estimates, i.e., $\text{var}(\Psi(\hat{Q}_n)) = \text{var}[\Psi(Q^*_{n,b})]$.

We emphasize that this TMLE is provided such that we do not need to re-estimate $\bar{Q}_n, g_n$. If $g_n \to g_0$ and $\bar{Q}_n \to \bar{Q}_0$, then this TMLE is asymptotically linear with IF $D^*(q_0, g_0)$. This is conservative relative to the variance of the actual TMLE that is estimated with g_n fitted on the data, when g_n is consistent.

5 Simulations

Simulation studies presented in this section investigate the performance of the two proposed variance estimators: (1) the "robust" approach based on a TMLE of the variance itself (3) and (2) the computationally efficient bootstrap based on a modified TMLE (4). We further compare the performance of these estimators to the common approach to variance estimation based on the sample variance of the empirical EIF (using either the initial $\bar{Q}^d_t$ or targeted $\bar{Q}^{d,*}_{t,n}$ estimators of the q_0 factors of the likelihood). We consider two data-generating processes and corresponding target causal parameters: a point treatment setting and a longitudinal observational study setting with three time points (i.e., $K+1 = 3$) with time-dependent confounding. To evaluate the performance of the variance estimators, we first compare the mean of the variance estimates to the Monte Carlo variance of the point estimator; we also report the Monte Carlo variance of each variance estimator. Additionally, we present the 95% confidence interval coverage, Type I and Type II errors resulting from each variance estimation approach. All analyses were conducted on R version 3.1.1 [24]. Codes corresponding to the simulations have been uploaded to https://github.com/tranlm/tmleVariance.

5.1 Data-generating distribution P_0 and causal parameters

5.1.1 Point treatment setting

Consider a point treatment setting, such as patient enrollment into a care program, in which the treatment $A(0)$ is only assigned at a single time point. We are interested in determining whether the treatment of interest has an effect on the (binary) outcome on an additive scale. Our target parameter is therefore the difference of the mean outcomes under treatment and control, i.e., $\psi_{0,1} \equiv \mathbb{E}[Y_1 - Y_0]$. Under this setting, the simulated data were generated (such that we could observe the levels of positivity violations desired) as follows:

$$\begin{aligned}
W_1 &\sim N(0,1), \text{ bounded at } [-2,2],\\
W_2 &\sim \text{Ber}(\text{logit}^{-1}(-1)),\\
L_1(0) &\sim N(0.1 + 0.4W_1, 0.5^2),\\
L_2(0) &\sim N(-0.55 + 0.5W_1 + 0.75W_2, 0.5^2),\\
\bar{g}_{0,0}(\text{Pa}(A(0))) &= \text{logit}^{-1}(\beta_p - (\beta_p + 2.5)W_1 + 1.75W_2 + (\beta_p + 3.2)L_1(0) - 1.8L_2(0) + 0.8L_1(0)L_2(0)),\\
\bar{Q}_{0,1}(\text{Pa}(Y)) &= \text{logit}^{-1}\Big(-0.5 + 1.2W_1 - 2.4W_2 - 1.8L_1(0) - 1.6L_2(0) + L_1(0)L_2(0) - \beta_{\psi_0}A(0)\Big),
\end{aligned}$$

with a positivity-associated parameter β_p ranging from –2 (minor positivity violations) to 0 (strong practical positivity violations). In estimation of $g_{0:t}$, we impose a truncation of 0.001. For the treatment effect-

associated parameter β_{ψ_0}, we consider values ranging from 0 (no treatment effect) to 1 (strong treatment effect). Here, $L_1(0)$ and $L_2(0)$ are not the time-dependent confounders and are therefore considered baseline covariates along with (W_1, W_2), which affect both the treatment and the outcome.

Under these settings, the true parameter values ψ_0 were achieved by generating 10^8 observations under the counterfactual distribution for each β_{ψ_0} considered. Simulation results were obtained for 1,000 simulations of size $n = 500$. Within each simulation, the bootstrap estimates of variance were formed from $B = 1{,}000$ iterations.

5.1.2 Longitudinal treatment setting

For the longitudinal setting, we considered a treatment $A(t)$ and outcome $L_3(t)$, which are each allowed to vary over time as a counting process (e.g., the treatment variable could be enrollment into a health program, while the outcome variable could be survival up to that time point). That is, if $A(t) = 1$, then we have that $\underline{A}(t) = 1$, where $\underline{A}(t) = (A(t), A(t+1), \ldots, A(K))$. Similarly, if $L_3(t) = 1$, then $\underline{L_3}(t) = 1$.

We are (again) interested in whether the treatment of interest has an effect on the outcome at the final time point $t^* = 3$ on an additive scale. Thus, our target parameter is the difference of the mean outcomes under treatment and control at this final time point, i.e., $\psi_{0,3} = \mathbb{E}\,[Y_1(t^*) - Y_0(t^*)]$, where $Y(t^*) = L_3(3)$. Under this setting, data for the first time point were generated in the same manner as the point treatment setting in Section 5.1.1. For the remaining two time points, the data were generated (again such that we could observe the levels of positivity violations desired) conditional on survival (i.e., $L_3(t^-) = 0$) as follows:

$$
\begin{aligned}
L_1(t) &\sim N\Big(0.1 + 0.4W_1,\, 0.5^2 + 0.6L_1(t^-) - 0.7L_2(t^-) + 0.45\beta_{\psi_0}A(t^-)\Big),\\
L_2(t) &\sim N\Big(-0.55 + 0.5W_1 + 0.75W_2 + 0.1L_1(t^-) + 0.3L_2(t^-) + 0.75\beta_{\psi_0}A(t^-),\, 0.5^2\Big),\\
\bar{g}_{0,t}(\mathrm{Pa}(A(t))) &= \mathrm{logit}^{-1}(\beta_p - (\beta_p + 2.5)W_1 + 1.75W_2 + (\beta_p + 3.2)L_1(t) - 1.8L_2(t) + 0.8L_1(t)L_2(t)),\\
\bar{Q}_{0,t}(\mathrm{Pa}(L_3(t))) &= \mathrm{logit}^{-1}\Big(-0.5 + 1.2W_1 - 2.4W_2 - 1.8L_1(t^-) - 1.6L_2(t^-) + L_1(t^-)L_2(t^-) - \beta_{\psi_0}A(t^-)\Big).
\end{aligned}
$$

Similar to the point treatment setting, the treatment effect-associated parameter β_{ψ_0} also ranged from 0 to 1. We note, however, that the positivity issues faced in this scenario will be even more severe due to the higher number of combinations of treatment over time, which result in a larger number of truncations. Figure 1 shows the proportion of observations with truncated $g_{0:2}$ as a function of β_p at a null effect, i.e., $\beta_{\psi_0} = 0$.

5.2 Estimators

5.2.1 TMLE specification

Any submodel and loss function for which its loss-function-specific score

$$
\frac{\partial}{\partial \varepsilon} L(P(\varepsilon))\bigg|_{\varepsilon=0}
$$

spans $D^*(P_0)$ can be chosen in TMLE for both the estimation of the mean outcome $\mathbb{E}[Y_a]$ and the variance of the EIF σ^2. As the corresponding estimators solve the estimating equation corresponding to the EIF, they will all be asymptotically equivalent and thus all be asymptotically efficient. That is, no difference will be seen between TMLEs defined using alternative submodels as the sample size grows to infinity. In the TMLE presented by van der Laan and Gruber [9], these submodels are used in the targeting step for each $\bar{Q}_t$ using a loss $L(\bar{Q}_t)$ that is indexed by $\bar{Q}_{t+1}$. Specifically, for the targeting step, we need a loss and submodel with clever covariate such that the score given solves a desired component of the EIF $D_t^*(P_0)$.

In finite samples, however, TMLEs defined using various submodels can have varying performance. For example, under increasing levels of positivity violations, the use of linear submodels that use $H_t(g)$ as a covariate can have higher variance due to observations with low probabilities of treatment acting as high leverage points, which result in highly influential points for the estimation of the submodel parameter ε.

Recall that the catalyst for this work was the anti-conservative estimates of estimator variance resulting from the use of the empirical EIF variance. We therefore wish to establish a robust estimator of the variance of estimators that solve the EIF, particularly under violations or near violations of positivity. In other words, we desire a variance estimator that will asymptotically converge to the true variance of the estimator, but also simultaneously act on the conservative side in finite samples. Keeping this in mind, we used two submodel and loss function combinations for our simulations, in line with the recommendations above. For point estimation of the target parameter and the robust estimator of the EIF variance (3), we used submodels that define $H_t(g)$ and $H_m^{d,t}(g)$ to be observational weights such that

$$\operatorname{logit}\bar{Q}(\varepsilon) = \operatorname{logit}\bar{Q} + \varepsilon\mathbb{I}(\bar{A}(t) = \bar{a}(t)),$$

acknowledging our slight abuse of notation. Alternatively, in our bootstrap approach at estimating the TMLE variance, we define a clever covariate using $H_t(g)$ such that

$$\operatorname{logit}\bar{Q}(\varepsilon) = \operatorname{logit}\bar{Q} + \varepsilon H_t(g).$$

Both submodels use, as loss function, the negative log-likelihood loss.

5.2.2 Nuisance parameter estimators

To estimate our nuisance parameters $\bar{Q}_0$ and g_0 in the point treatment setting, we fit linear models and estimate the coefficients using maximum-likelihood estimation. For example, we fit the treatment mechanism by using logistic regression with the following (correct) specification:

$$\operatorname{logit}^{-1}(\beta_0 - \beta_1 W_1 + \beta_2 W_2 + \beta_3 L_1(0) - \beta_4 L_2(0) + \beta_5 L_1(0)L_2(0)).$$

For the longitudinal setting, we (mostly) followed the same approach as in the point treatment with two exceptions:

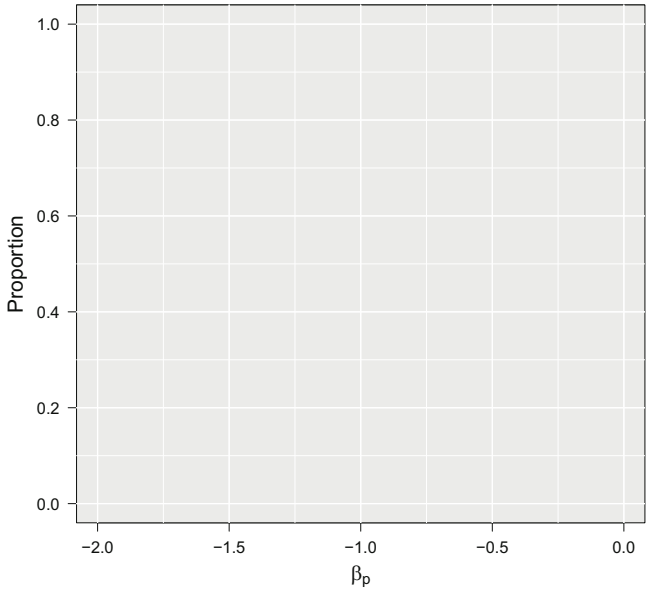

Figure 1: Proportion of observations with $g_{0:2}$ truncated at each β_p.

(1) For estimating the treatment mechanism, we pool the data across the time points (after conditioning on not yet having received treatment). This provides greater data support for estimating the probability of treatment across all three time points, resulting in higher precision for the estimates of the coefficients.
(2) We continue to use the same specification (across all time points) for estimating $\bar{Q}_0$ (e.g., for time 2, we use $\beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 L_1(2) + \beta_4 L_2(2) + \beta_5 L_1(2) * L_2(2) + \beta_6 A(2)$). While this is not the true model for $\bar{Q}_0$, we still have consistency in our estimate of $\psi_{0,1}$ due to our consistent estimation of g_0.

Of note, in both the single time point and longitudinal settings, we use a correctly specified parametric model for the treatment mechanism. For the single time point setting, we use a correctly specified parametric model for the outcome regression, whereas in the longitudinal setting, we use a mis-specified parametric model. This allows us to investigate how our variance estimators perform as we deviate from correct specification of the outcome regressions. This is further motivated by the fact that the double robust estimators that we study remain asymptotically linear in this setting, while the variance estimator becomes conservative (because it targets an IF with higher variance).

5.3 Simulation results

5.3.1 Point treatment results

Figure 2 shows the Monte Carlo variance under no treatment effect ($\beta_{\psi_0} = 0$) for the TMLE point estimator, along with the mean of the variance estimates from each estimation approach. At the lower end of β_p, where positivity violations are minor, the observed estimator variance is low. For example, at $\beta_p = -2$, the Monte Carlo variance was 8.068×10^{-4}. As β_p increased, introducing higher levels of positivity violations, the estimator variance increases as expected.

Regarding the mean of variance estimator (and thus its bias), all four approaches were similar at low values of β_p. For example, at $\beta_p = -2$, the mean of the estimates was 7.957×10^{-4}, 7.563×10^{-4}, and

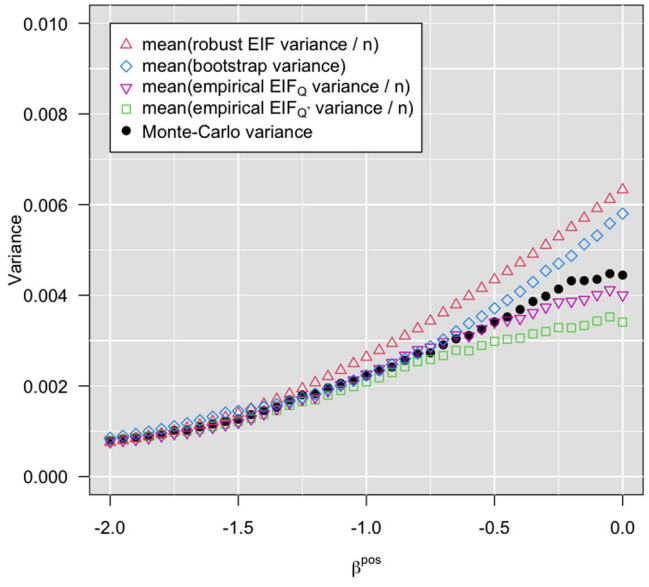

Figure 2: Mean of variance estimates under no treatment effect ($\beta_{\psi_0} = 0$) at each positivity (β_p) value under the point treatment setting, overlaid with the estimator's Monte Carlo variance.

8.545×10^{-4} for the empirical EIF ($\bar{Q}_t^d$), robust, and bootstrapped-based approaches, respectively, compared to the estimator's Monte Carlo variance of 7.977×10^{-4}. As β_p increased, the empirical EIF approaches underestimated the variance on average (although to a lesser extent when the targeted, as compared to initial, estimate of the q_0 factors was used). In contrast, the bootstrap and robust EIF approaches resulted in slightly conservative estimates.

Figure 3 shows the Monte Carlo variance for each method used for estimating the variance. Lower values in this figure can be interpreted as coming from a variance estimator with higher precision. From it, we can see that the empirical EIF approach to estimating variance has the highest variance at severe levels of positivity violation. Additionally, the use of $\bar{Q}_t^d$ results in noticeably higher variance than $\bar{Q}_{t,n}^{d,*}$. Conversely, the robust approach of estimating variance maintains its low level of variability across all positivity settings. The bootstrap approach tended to result in variance that was in between, though converged closer to the empirical EIF approach at high levels of positivity violations.

We evaluated 95% confidence interval coverage for the TMLE estimator of $\mathbb{E}Y_d$ under the four approaches to variance estimation. Figure 4 shows a heat map overlaid with a single contour line (at 0.95) of the resulting coverage estimates (i.e., the observed proportion of times the true parameters were captured by the confidence intervals) over the different combinations of β_{ψ_0} and β_p. Additionally, we estimated the power to reject the null hypothesis (at a level of 0.05) corresponding to each variance estimation approach under the range of treatment effect sizes and degrees of positivity violation considered above. Figure 5 shows a heat map overlaid with a single contour line (at 0.05) of the resulting power estimates. Results at $\beta_{\psi_0} = 0$ can be interpreted as Type I errors, as they inform us of the times that the null hypothesis of no treatment effect is incorrectly rejected.

The empirical EIF approaches consistently demonstrated a slight under-coverage of the true parameter values (even at low levels of positivity violations), which increased in severity with the severity of positivity violations. For example, coverage ranged from 0.945 at $\beta_p = -2$ to 0.872 at $\beta_p = 0$. In contrast, the robust EIF approach consistently resulted in coverage at around 0.95–0.96 at low values of β_p and *increased* with β_p, consistent with expectation for the overestimation of the variance under increasing positivity by this approach. For example, at $\beta_p = -0.7$, coverage remained at 0.98 at all values of β_{ψ_0}. At $\beta_p \geq -0.1$, the observed coverage was almost always greater than or equal to 0.99 at all values of β_{ψ_0}. As the figure only

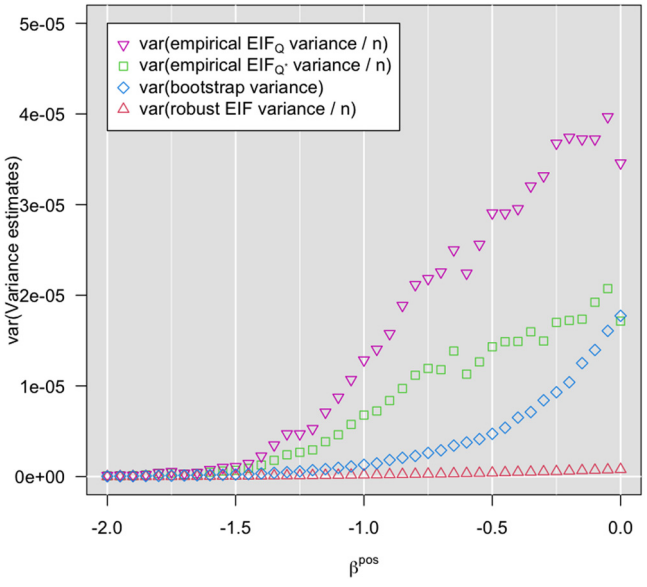

Figure 3: Monte Carlo variance of variance estimators under no treatment effect ($\beta_{\psi_0} = 0$) at each positivity (β_p) value under the point treatment setting.

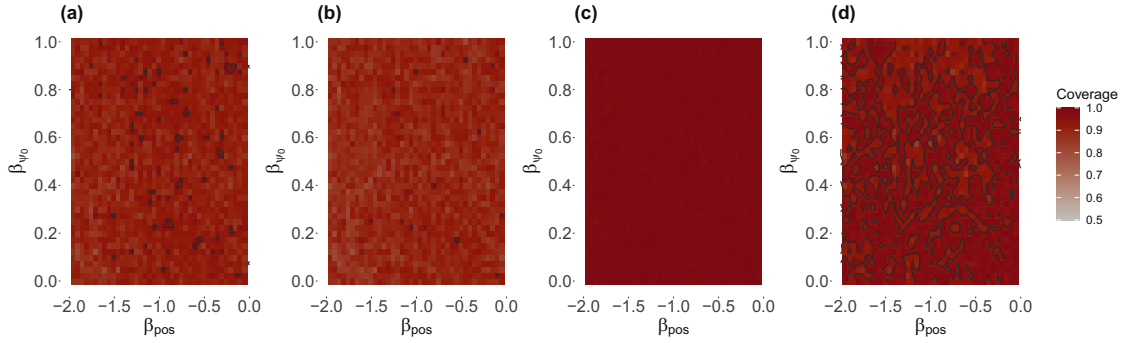

Figure 4: Simulated 95% coverage for each variance estimation approach for the TMLE point estimator under various treatment (β_{ψ_0}) and positivity (β_p) values under the point treatment setting. Contour line corresponds to 95%. (a) Empirical EIF Q, (b) empirical EIF Q∗, (c) robust EIF, and (d) bootstrap.

contains a single contour corresponding to 95%, neither the empirical EIF (consistently under 0.95) nor the robust EIF approach (consistently over 0.95) contains the lines. The bootstrap-based coverage shown in Figure 4(d) varied the least, with coverage consistently between 0.95 and 0.97 irrespective of the treatment effect (β_{ψ_0}) and positivity severity (β_p) considered.

Regarding the observed power (Figure 5), the empirical-EIF-based variance approach resulted in the highest power among all three variance estimation approaches when an effect was present. For example, at $\beta_{\psi_0} = 1$ and $\beta_p = -1$, the observed power was 0.71, 0.51, and 0.51 for the empirical-EIF (Q), robust-EIF, and bootstrap approaches, respectively. However, this gain came at a cost of higher Type I error, which became uncontrolled as positivity violations increased (i.e., with an increase in β_p). For example, at $\beta_p = -2$ an observed 5.2% of the simulations incorrectly rejected the null hypothesis. This proportion increased to as high as 12% at $\beta_p = 0$. Alternatively, the robust EIF estimation approach resulted in nominal to low Type I errors (i.e., between 0 and 5.1%). The bootstrap approach resulted in the best performance overall of the estimators considered, with higher power than the robust EIF approach when an effect was present while simultaneously retaining appropriate control of the Type I error at all levels of β_p when no effect was present. For example, at $\beta_p = 0$, only 4.6% of the simulations incorrectly rejected the null hypothesis.

To further investigate the performance, we evaluated (i) the sampling distribution of the point estimator and (ii) the distribution of the variance estimator, conditioning on the data-generating process with the highest level of positivity violations ($\beta_p = 0$) and no treatment effect ($\beta_{\psi_0} = 0$). The results (Figure 6(a)) show that the distribution of the point estimators is approximately normal, despite having somewhat heavier tails, suggesting that lack of normality of the estimator itself is not the primary driver of the loss of coverage observed. Instead, results suggest that under-coverage is likely due primarily to the highly

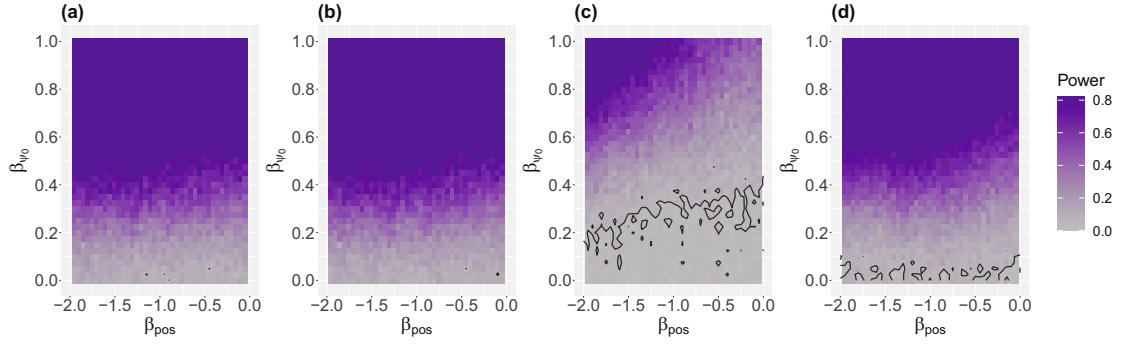

Figure 5: Simulated power for each variance estimation approach for the TMLE point estimator under various treatment (β_{ψ_0}) and positivity (β_p) values under the point treatment setting. Contour line corresponds to 5%. (a) Empirical EIF Q, (b) empirical EIF $Q*$, (c) robust EIF, and (d) bootstrap.

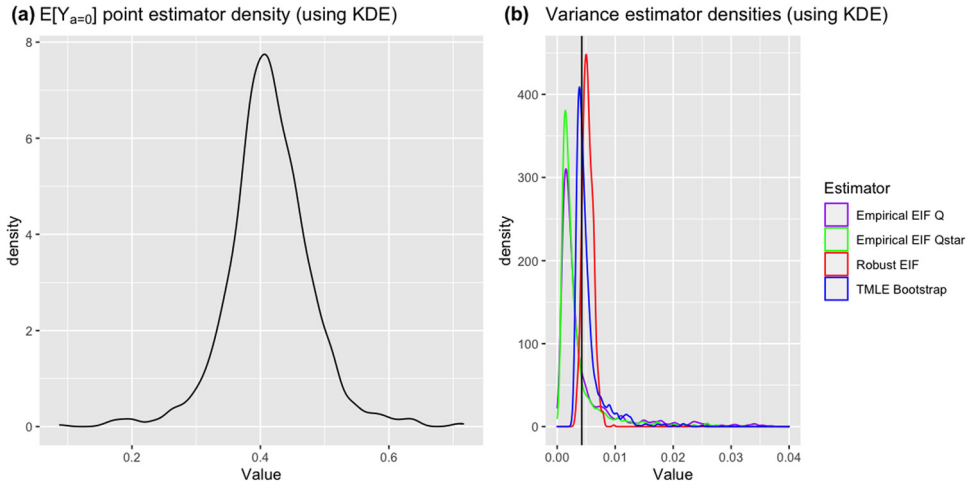

Figure 6: Under no treatment effect ($\beta_{\psi_0} = 0$) at the most severe positivity setting ($\beta_p = 0$). (a) Sampling distribution of TMLE point estimates and (b) sampling distribution of variance estimates with overlaid Monte Carlo variance (black vertical line).

right-skewed distribution of the empirical EIF variance estimators. The distribution of the empirical EIF approach of estimating variance is noticeably skewed (Figure 6(b)), with most of the probability mass well under the Monte Carlo variances. Conversely, the bootstrap variance estimator is less skewed, with a distribution that is more closely centered around the observed Monte Carlo variance. The robust EIF variance estimator is also noticeably less skewed, though it tends to overestimate the variance.

5.3.2 Longitudinal treatment results

Results for the longitudinal setting are similar to the point treatment setting. Similar to the point treatment setting, Figure 7 shows the mean of the variance estimates under each approach, overlaid with the Monte Carlo variance of the intervention-specific mean outcome point estimators. The same trend over the

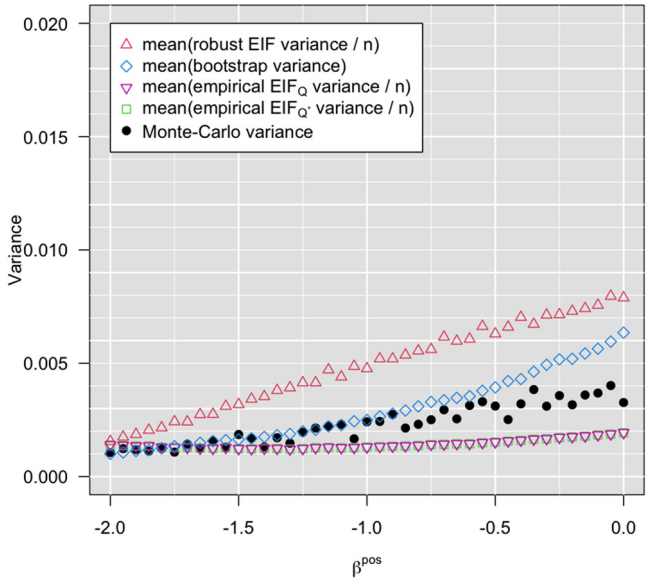

Figure 7: Mean of variance estimates for each estimator under no treatment effect ($\beta_{\psi_0} = 0$) at each positivity (β_p) value under the longitudinal treatment setting, overlaid with the estimator's Monte Carlo variance.

different levels of positivity was seen as in Figure 3, with the variance increasing with increasing magnitude of positivity violations. The empirical EIF approach also performed well here at low levels of β_p. At high values of β_p, the approach more noticeably underestimates the variance of the intervention-specific mean outcome point estimator. Consistent with the point treatment setting, the mean of the robust EIF variance estimator overestimated the variance. The bootstrap approach resulted in variance estimates that were slightly conservative on average, though were still very similar to the Monte Carlo variance estimates.

Figure 8 shows the coverage corresponding to each variance estimation approach for the TMLE point estimator of the intervention-specific mean outcome. Coverage for the empirical EIF approaches was consistently anti-conservative, even at low levels of positivity violations. As in the point treatment setting, this was true despite the apparent low bias of this variance estimator, likely due again to its heavily skewed distribution. For example, at a null effect (i.e., β_{ψ_0}), the observed coverage was 0.91 at $\beta_p = -2$ and 0.87 at $\beta_p = 0$ for the empirical EIF Q. For the robust EIF approach, coverage remained conservative. This became as high as 1.00 (i.e., all simulated confidence intervals captured the true parameter value) at higher levels of positivity issues. For the bootstrap approach, close to nominal coverage was seen, even for higher levels of positivity violations. For example, under a null effect, a coverage of 0.97 was observed at $\beta_p = -2$ and 0.95 at $\beta_p = 0$.

Results for the Type I error and power were also similar to the point treatment setting. When there was an effect, the empirical EIF approach resulted in the highest power. At $\beta_{\psi_0} = 1$ and $\beta_p = -2$, we observed a power of 0.99. However, the Type I error was also uncontrolled here, becoming as high as 0.14 at $\beta_p = 0$. In contrast, the robust EIF approach resulted in overly conservative Type I error rates, particularly in the context of greater positivity violation, and thus, the power for this approach when an effect was present was the lowest. For example, for a treatment effect size of $\beta_{\psi_0} = 1$, we observed a power ranging from 0.94 at $\beta_p = -2$ to 0.47 at $\beta_p = 0$. The bootstrap approach resulted in generally well-controlled Type I error rates, with observed values ranging between 0.02 and 0.08. Power was also higher than the robust EIF approach across all values of β_{ψ_0} and β_p. For a treatment effect size of $\beta_{\psi_0} = 1$, we observed a power ranging from 0.99 at $\beta_p = -2$ to 0.93 at $\beta_p = 0$ for the bootstrap approach. Compared with the robust EIF approach, this is up to a twofold increase in power (Figure 9).

6 Discussion

The goal of this study was to establish a consistent and robust approach for estimating the variance of asymptotically efficient estimators such as TMLE, estimating equations, and one-step estimators which, in contrast to the common approach based on the empirical variance of the estimated EIF, does not provide

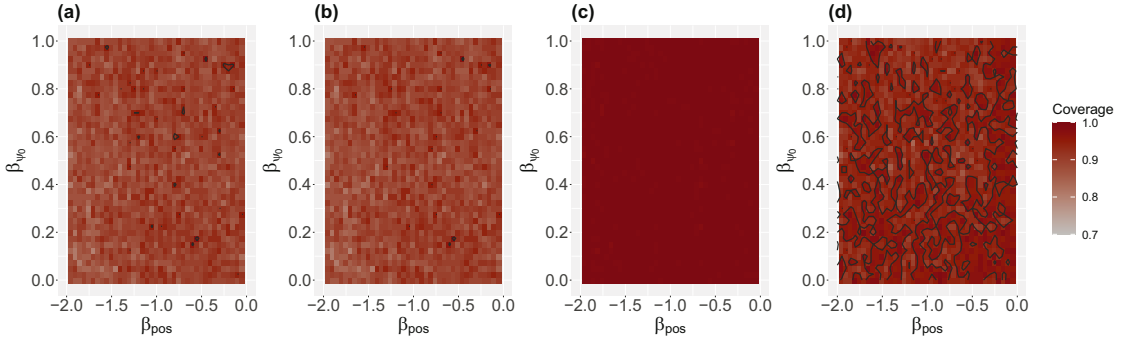

Figure 8: Simulated 95% confidence interval coverage for each variance estimation approach for the TMLE point estimator under various treatment (β_{ψ_0}) and positivity (β_p) values under the longitudinal treatment setting. Contour line corresponds to 95%. (a) Empirical EIF Q, (b) empirical EIF $Q*$, (c) robust EIF, and (d) bootstrap.

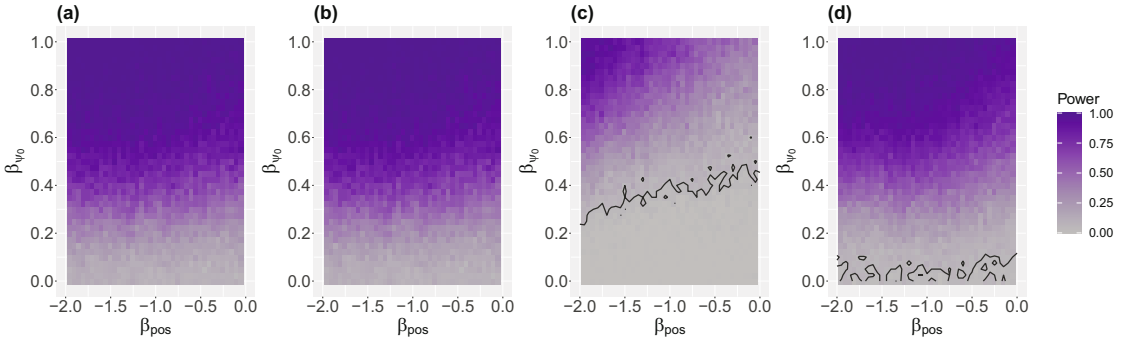

Figure 9: Simulated power for each variance estimation approach for the TMLE estimator under various treatment (β_{ψ_0}) and positivity (β_p) values under the longitudinal treatment setting. Contour line corresponds to 5%. (a) Empirical EIF Q, (b) empirical EIF Q^*, (c) robust EIF, and (d) bootstrap.

anti-conservative inference when confronted with positivity violations. We have presented two such approaches for estimating this variance: (1) a robust approach that directly targets the asymptotic variance of the EIF and (2) a bootstrap approach based on fitting the initial density of the data once, followed by a non-parametric bootstrap of the targeting step. In simulations, the variance of the point estimator increased with increases in positivity violations as expected. A common trend was observed, with the empirical EIF approach of variance estimation providing anti-conservative confidence interval coverage (as previously reported). In contrast, the robust EIF approach described here resulted in increasingly conservative coverage as the degree of positivity violation increased. The bootstrap-based approach provided closer to nominal 95% confidence interval coverage and Type I error control (with attendant power gains relative to the robust estimator) in the face of extreme positivity violations, both in the point treatment and in the longitudinal setting.

While the EIF can raise a red flag for lack of identifiability, for substitution estimators such as TMLE, we suggest that it is overly conservative for constructing valid confidence intervals and tests in finite sample in the face of substantial positivity violations. Previous work [12] suggested the use of the parametric bootstrap as a diagnostic tool for sparsity-bias in the point treatment setting. The approach can become cumbersome, as the analyst would need to refit the whole likelihood for each iteration of the bootstrap, and computationally intensive, particularly in longitudinal treatment settings. Our robust EIF approach is able to avoid estimating the whole likelihood by targeting the required means under the post-intervention distribution defined by the longitudinal g-computation formula directly. Even if we use an actual TMLE of P_0^d, our analytic estimate of the variance is still much less computer-intensive than the parametric bootstrap method, in particular, in view that one would need to run many replicate samples of the data set in order to pick up the observations that correspond with a rare treatment and thus contribute large numbers to the variance expression. Our presented bootstrap approach, while more computationally intensive than the robust EIF approach, is also superior to the earlier proposed approach, in that we do not have to refit the entire likelihood under each iteration. This also significantly reduces the computational resources required to obtain estimates of the target parameter, particularly if computationally intensive non-parametric machine learning algorithms are used to estimate these densities. Therefore, we believe that the proposed analytic method will be (at least, practically) superior to the earlier proposed parametric bootstrap method.

In order to highlight challenges to variance estimation arising from positivity violations, and the extent to which these are addressed by the alternative variance estimators proposed, this article has focused on the setting in which the treatment mechanism is estimated using a correctly specified parametric model, while the outcome regressions are estimated using either a correct or mis-specified parametric model. In practice, outside of randomized trial settings, non-parametric machine learning approaches are typically required to ensure asymptotic linearity of the estimator (see, e.g., Klaassen [25], van der Laan and Robins [26]). In this setting, valid inference also requires adequate convergence rate of the second-order remainder. In particular, the relatively slow convergence of machine learning estimators poses a serious challenge to finite

sample inference due to the non-negligible contribution of the second-order remainder. One possible response to this challenge is to incorporate data-adaptive nuisance parameter estimation into our proposed non-parametric bootstrap, an approach shown to be promising in prior work [27], although not evaluated in the context of positivity violations. Future research is needed to evaluate the performance of this approach under the joint challenges of positivity violations and slower machine-learning-based convergence rates.

Further refinements can be applied in an attempt to obtain variance estimates with an even smaller finite sample bias. One such approach is a convex combination of the variance estimators considered earlier. For example, we noted in supplementary analyses that taking

$$\hat{\alpha}_n \hat{\sigma}^2_{e\text{EIF},n} + (1 - \hat{\alpha}_n)\hat{\sigma}^2_{r\text{EIF},n}$$

had good performance, where $\hat{\sigma}^2_{e\text{EIF},n}$ is the variance estimate using the empirical EIF approach, $\hat{\sigma}^2_{r\text{EIF},n}$ is the variance estimate using the robust EIF approach, and $\hat{\alpha}_n = |\hat{\sigma}^2_{r\text{EIF},n} - \hat{\sigma}^2_{e\text{EIF},n}| / (\hat{\sigma}^2_{r\text{EIF},n} + \hat{\sigma}^2_{e\text{EIF},n})$. We note, however, that such an approach is somewhat *ad hoc* and may lead to varying results in other simulations or distributions. We therefore chose not to present the results here.

A potential limitation of the robust approach at estimating the variance involves the conditions for asymptotic linearity to be met. Note, however, that we always require that our estimator (of the parameter) be asymptotically linear. Thus, this is actually a limitation of our parameter estimator. Given that we have an asymptotically linear estimator, we want a good estimator of its variance. If we correctly specify both the outcome and treatment models, then the variance is equal to the variance of the EIF and we can proceed with applying our robust method of estimating variance knowing that we are targeting the correct variance. However, the variance of the EIF is not consistent for the variance of the effect estimator if either of the outcome or treatment models are not correctly specified.

Furthermore, it is also required that $\bar{Q}_0^{d,\sigma_t^2}$ be estimated both consistently and at a fast enough rate. We limited the computational complexity in our simulations by using simpler parametric models to estimate $\bar{Q}_0^{d,\sigma_t^2}$, though a more non-parametric approach such as Super Learning could have been applied. This approach can become computationally expensive if there are many time points. In this regard, the bootstrap approach is superior as it does not require the additional estimation of $\bar{Q}_0^{d,\sigma_t^2}$.

It would be of interest to further evaluate not only the practical performance of these variance estimation approaches in future studies, but also the application of the approaches to other parameters. Appendix A derives the general approach for working marginal structural models. Further research into the practical performance of this approach is needed for this setting. These variance estimation approaches can also be used for more complex parameters, such as mean outcomes under dynamic regimes, stochastic interventions, or marginal structural working models. Future research could also develop a collaborative TMLE [28] or cross-validated [29] based approach at robustly estimating the EIF variance.

Funding information: This work was supported by the Doris Duke Clinical Scientist Development Award (NIH-NIAID U01AI069911), The National Institute of Allergy and Infectious Diseases of the National Institutes of Health (U01AI069911), The President's Emergency Plan for AIDS Relief (PEPFAR) (AID-623-A-12-0001), and NIH (R01AI074345).

Conflict of interest: Prof. Maya Petersen and Prof. Mark J. van der Laan are Editors of the Journal of Causal Inference but were not involved in the review process of this article.

References

[1] Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. J Amer Stat Assoc. 1952;47(260):663–85.

[2] Robins JM. Marginal structural models. 1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science. 1998. p. 1–10. http://link.springer.com/chapter/10.1007/978-1-4419-9782-1_9.
[3] Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell NP, Dietz K, Farewell VT, editors. AIDS epidemiology. Boston: Birkhäuser; 1992. p. 297–331.
[4] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Amer Stat Assoc. 1994;89(427):846–67.
[5] Robins JM, Rotnitzky A, van der Laan MJ. Discussion of "On profile likelihood" by Murphy and van der Vaart. J Amer Stat Assoc. 2000;95(450):477–82.
[6] Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. In: Proceedings of the American Statistical Association Section on Bayesian Statistical Science; 2000. p. 6–10.
[7] Robins JM, Rotnitzky A. Comment on the Bickel and Kwon article, "Inference for semiparametric models: some questions and an answer". Statistica Sinica. 2001;11(4):920–36.
[8] Rotnitzky A, Robins J. Inverse probability weighted estimation in survival analysis. In Wiley StatsRef: Statistics Reference Online (N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels editors); 2014.
[9] van der Laan MJ, Gruber S. Targeted minimum loss based estimation of an intervention specific mean outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 290.
[10] Hampel FR. The influence curve and its role in robust estimation. J Amer Stat Assoc. 1974;69(346):383–93. http://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10482962.
[11] Bradley E.. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Stanford, CA: Stanford University; 1981. p. 3. https://statistics.stanford.edu/sites/default/files/EFSNSF156.pdf.
[12] Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Meth Med Res. 2012 Feb;21(1):31–54.
[13] Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan M. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models HHS public access. J Causal Inference. 2014;2(2):147–85.
[14] Schwab J, Lendle S, Petersen M, van der Laan MJ. LTMLE: longitudinal targeted maximum likelihood estimation. R package version 0.9.3-1. https://CRAN.R-project.org/package=ltmle; 2014.
[15] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005 Dec;61(4):962–73. http://www.ncbi.nlm.nih.gov/pubmed/16401269.
[16] Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. Math Modell. 1986;7:1393–512.
[17] Robins JM. Semi-parametric efficiency in multivariate regression models with missing data. J Amer Stat Assoc. 1995;(90):122–29.
[18] Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry D, editors. Statistical Models in Epidemiology, the Environment, and Clinical Trials. The IMA Volumes in Mathematics and its Applications, vol 116. Springer, New York, NY. 1999. https://doi.org/10.1007/978-1-4612-1284-3_2.
[19] van der Laan MJ, Rubin D. Targeted maximum likelihood learning. UC Berkeley Division of Biostatistics Working Paper Series. 2006;(213):1–87.
[20] Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007 Jan;22(4):523–39.
[21] Robins J, Sued M, Lei-gomez Q, Rotnitzky A. Comment: Performance of double-Robust estimators when "Inverse Probability" weights are highly variable. Stat Sci. 2007;22(4):544–59.
[22] Rotnitzky A, Lei Q, Sued M, Robins JM. Improved double-robust estimation in missing data and causal inference models. Biometrika. 2012;99(2):1–18.
[23] Gruber S, van der Laan MJ. Targeted maximum likelihood estimation: a gentle introduction. UC Berkeley Division of Biostatistics Working Paper Series; 2009. p. 252.
[24] R CoreTeam. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2014. https://www.R-project.org/.
[25] Klaassen CAJ. Consistent estimation of the influence function of locally asymptotically linear estimators. Ann Stat. 1987;15(4):1548–62.
[26] van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. New York: Springer; 2003.
[27] Cai W, van der Laan MJ. Nonparametric bootstrap inference for the targeted highly adaptive LASSO estimator. arXiv:1905.10299; 2019.
[28] van der Laan MJ, Gruber S. Collaborative Double Robust Targeted Maximum Likelihood estimation. Int J Biostat. 2010;6(1):Article 17.
[29] Zheng W, van der Laan M. Asymptotic theory for cross-validated targeted maximum likelihood estimation. UC Berkeley Division of Biostatistics Working Paper Series; 2010. p. 273. http://biostats.bepress.com/ucbbiostat/paper273/.

Appendix

A Marginal structural working model variance

A.1 TMLE of σ^2_{K+1} for marginal structural working models

For the general working logistic marginal structural model (MSM) $\Theta \equiv \{m_\beta : \beta\}$ from Petersen et al. [12], we have that the component of the EIF corresponding with the last time point $K+1$ equals

$$\begin{aligned} D^*_{K+1}(P) &= \sum_{d \in \mathcal{D}} h_1(d, K+1) \frac{\mathbb{I}(\bar{A}(K) = d(\bar{L}(K)))}{g_{0:K}(\bar{A}(K), \bar{L}(K))} (Y - \bar{Q}_{K+1}(\bar{A}(K), \bar{L}(K))) \\ &= C_{K+1}(P)(\bar{A}, \bar{L})(Y - \bar{Q}_{K+1}), \end{aligned}$$

where, for some user-defined weight function $h(d, K+1)$,

$$\begin{aligned} C_{K+1}(P)(\bar{A}, \bar{L}) &= \sum_{d \in \mathcal{D}} h_1(d, K+1) \frac{\mathbb{I}(\bar{A}(K) = d(\bar{L}(K)))}{g_{0:K}(\bar{A}, \bar{L})}, \\ h_1(d, K+1) &= h(d, K+1) \frac{\frac{\partial}{\partial \beta} m_\beta(d, K+1)}{m_\beta(1 - m_\beta)}. \end{aligned}$$

Note that we want to obtain a representation of the variance of this component D^*_{K+1} so that we can use a semi-substitution estimator of this part of the variance of the EIF, hopefully, thereby obtaining a variance estimator that is more accurate under violations of practical positivity, and a variance estimator that can be used as a red flag for lack of identifiability. This variance can thus be written as:

$$\begin{aligned} \sigma^2_{K+1} &= \mathbb{E}[C^2(Y - \bar{Q}_{K+1})^2] \\ &= \mathbb{E}[C^2 \bar{Q}_{K+1}(1 - \bar{Q}_{K+1})] \\ &= \mathbb{E}\left[\left(\sum_{d \in \mathcal{D}} h_1(d, K+1) \mathbb{I}(\bar{A} = d(\bar{L})) \right)^2 \frac{\bar{Q}_{K+1}(1 - \bar{Q}_{K+1})}{g^2_{0:K}}(O) \right] \\ &= \mathbb{E}\left[\left(\sum_{d_1, d_2 \in \mathcal{D}} h_1(d_1, K+1) h_1(d_2, K+1) \mathbb{I}(\bar{A} = d_1(\bar{L})) \mathbb{I}(\bar{A} = d_2(\bar{L})) \right) \frac{\bar{Q}(1 - \bar{Q})}{g^2_{0:K}}(O) \right] \\ &= \sum_{d_1, d_2} h_1(d_1, K+1) h_1(d_2, K+1) \mathbb{E}\left[\mathbb{I}(\bar{A} = d_1(\bar{L})) \mathbb{I}(\bar{A} = d_2(\bar{L})) \frac{\bar{Q}(1 - \bar{Q})}{g^2_{0:K}}(O) \right]. \end{aligned}$$

The latter expectation equals

$$\begin{aligned} &\int_{\bar{L}} \mathbb{I}(d_1(\bar{L}) = d_2(\bar{L})) \prod_{t=0}^{K+1} q(L_t | \bar{A}(t^-) = d_1(\bar{L}(t^-)), \bar{L}(t^-)) \frac{\bar{Q}(1 - \bar{Q})}{g_{0:t}(d_1(\bar{L}), \bar{L})} \\ &\quad = \mathbb{E}_{P_0^{d_1}}\left[\mathbb{I}\left(d_1\left(\bar{L}_{d_1}\right) = d_2\left(\bar{L}_{d_1}\right)\right) \frac{\bar{Q}(1 - \bar{Q})}{g_{0:K}}\left(d_1\left(\bar{L}_{d_1}\right), \bar{L}_{d_1}\right) \right]. \end{aligned}$$

This yields the following expression:

$$\begin{aligned} \sigma^2_{K+1} &= \sum_{d_1, d_2 \in \mathcal{D}} h_1(d_1, K+1) h_1(d_2, K+1) \mathbb{E}\left[\mathbb{I}\left(d_1\left(\bar{L}_{d_1}\right) = d_2\left(\bar{L}_{d_1}\right)\right) \frac{\bar{Q}(1 - \bar{Q})}{g_{0:K}}\left(d_1\left(\bar{L}_{d_1}\right), \bar{L}_{d_1}\right) \right] \\ &= \sum_{d_1 \in \mathcal{D}} h_1(d_1, K+1) \mathbb{E}\left[\left(\sum_{d_2 \in \mathcal{D}} h_1(d_2, K+1) \mathbb{I}\left(d_1\left(\bar{L}_{d_1}\right) = d_2\left(\bar{L}_{d_1}\right)\right) \right) \frac{\bar{Q}(1 - \bar{Q})}{g_{0:K}}\left(d_1\left(\bar{L}_{d_1}\right), \bar{L}_{d_1}\right) \right] \\ &= \sum_{d_1 \in \mathcal{D}} h_1(d_1, K+1) \mathbb{E} Z_{d_1}(d_1, K+1), \end{aligned}$$

where

$$Z(d_1, K+1) = \left(\sum_{d_2 \in \mathcal{D}} h_1(d_2, K+1) \mathbb{I}(d_1(\bar{L}(K)) = d_2(\bar{L}(K))) \right) \frac{\bar{Q}(1-\bar{Q})}{g_{0:K}(d(\bar{L}(K)), \bar{L}(K))},$$

so that the counterfactual of $Z(d_1, K+1)$ under intervention d_1 is given by:

$$Z_{d_1}(d_1, K+1) = \left(\sum_{d_2 \in \mathcal{D}} h_1(d_2, K+1) \mathbb{I}\left(d_1\left(\bar{L}_{d_1}(K)\right) = d_2\left(\bar{L}_{d_1}(K)\right)\right) \right) \frac{\bar{Q}(1-\bar{Q})}{g_{0:K}} \left(d_1\left(L_{d_1}(K)\right), L_{d_1}(K)\right).$$

A.1.1 Static regimens

In the special case that the class of dynamic regimens $\mathcal{D}$ consists only of static regimens $\bar{a}(K)$ so that there is only one and exactly one treatment such that $\bar{A}(K) = d(\bar{L}(K))$, then we have

$$Z(K+1) = h_1(\bar{A}, K+1) \frac{\bar{Q}(1-\bar{Q})}{g_{0:K}}(\bar{A}, \bar{L}),$$

so that

$$Z_d(K+1) = h_1(d, K+1) \frac{\bar{Q}(1-\bar{Q})}{g_{0:K}}(d(\bar{L}_d), \bar{L}_d).$$

In that case, we have

$$\sigma^2_{K+1} = \sum_{d \in \mathcal{D}} h_1(d, K+1)^2 \mathbb{E} Z_{1d}(K+1),$$

where $Z_1(K+1) = \bar{Q}(1-\bar{Q})/g_{0:K}(\bar{A}, \bar{L})$ and $Z_{1d}(K+1) = \bar{Q}(1-\bar{Q})/g_{0:K}(d(\bar{L}_d), \bar{L}_d)$.

It is important to note that in expressing our variance this way, we integrate out the indicator of treatment over $\bar{A}$, i.e., $\mathbb{I}(\bar{A} = d(\bar{L}))$. By getting rid of this indicator, we no longer rely as heavily on observations from subjects following treatment in estimating the variance of D^*_{K+1}. This particularly helps us when there is a lack of positivity, since subjects with low probabilities of desired treatment simply are not observed.

We have now shown that

$$\sigma^2_{K+1} = \sum_{d \in \mathcal{D}} h_1(d, K+1) \mathbb{E} Z_d(d, K+1),$$

where

$$Z(d_1, K+1) = \left\{ \sum_{d_2} h_1(d_2, K+1) \mathbb{I}(d_1(\bar{L}) = d_2(\bar{L})) \right\} \frac{\bar{Q}(1-\bar{Q})}{g_{0:K}}(d_1(\bar{L}), \bar{L}).$$

We can now define $Z(K+1)(\bar{A}, \bar{L}) = \sum_{d \in \mathcal{D}} h_1(d, K+1) \mathbb{I}(\bar{A} = d(\bar{L})) \frac{\bar{Q}(1-\bar{Q})}{g_{0:K}}(\bar{A}, \bar{L})$ (as a function of $\bar{A}, \bar{L}$) as a new outcome for our longitudinal data structure such that $Z_d(d, K+1) = Z(K+1)(d(\bar{L}_d), \bar{L}_d)$. Our variance σ^2_{K+1} is then represented as $\sum_{d \in \mathcal{D}} h_1(d, K+1) \mathbb{E} Z_d(d, K+1)$. Thus, if we redefine the longitudinal data as $(\bar{A}, \bar{L})$ with the final outcome of interest as $Z(K+1) = Z(K+1)(\bar{A}, \bar{L})$, and use the working MSM parameter $\mathbb{E} Z_d(K+1) = \beta_0$, with $\beta_0 = \arg\min_\beta \sum_{d \in \mathcal{D}} h_1(d, K+1)(\mathbb{E} Z_d(K+1) - \beta)^2$, then we have that

$$\beta_0 = \sum_{d \in \mathcal{D}} h_1(d, K+1) \mathbb{E} Z_d(K+1) / \sum_{d \in \mathcal{D}} h_1(d, K+1).$$

This demonstrates that we can obtain σ^2_{K+1} by simply multiplying β_0 by $\sum_{d \in \mathcal{D}} h_1(d, K+1)$, i.e.,

$$\sigma^2_{K+1} = \beta_0 \sum_d h_1(d, K+1).$$

We can therefore also estimate this variance component σ^2_{K+1} by computing the TMLE of the intercept β_0 in the working MSM for our newly defined outcome $Z(K+1)$ using weights $h_1(d, K+1)$ and then multiplying it against $\sum_{d\in\mathcal{D}} h_1(d, K+1)$.

A.2 TMLE of σ_t^2 for marginal structural working models

We now present how to obtain a TMLE of the variance of the tth component of the EIF, σ_t^2. For the general working MSM from Petersen et al. [12], we have that the component corresponding with the tth time point equals

$$\begin{aligned} D_t^*(P) &= \sum_{d\in\mathcal{D}} h_1(d,t) \frac{\mathbb{I}(\bar{A}(t^-) = d(\bar{L}(t^-)))}{g_{0:t^-}(\bar{A}(t^-), \bar{L}(t^-))} \left(\bar{Q}_{t^+}^d(\bar{A}(t), \bar{L}(t)) - \bar{Q}_t^d\right)(\bar{A}(t^-), \bar{L}(t^-)) \\ &= \sum_{d\in\mathcal{D}} C_t(P,d)\left(\bar{Q}_{t^+}^d - \bar{Q}_t^d\right). \end{aligned}$$

Similarly, we want to obtain a representation of the variance of this component so that we can use a semi-substitution estimator of this part of the variance of the EIF, hopefully, thereby obtaining a variance estimator that is more accurate under violations of practical positivity, and a variance estimator that can be used as a red flag for lack of identifiability. This variance σ_t^2 can thus be written as:

$$\sigma_t^2 = \sum_{d_1,d_2} h_1(d_1,t) h_1(d_2,t) \mathbb{E}\left[\mathbb{I}(\bar{A}(t^-) = d_1)\mathbb{I}(\bar{A}(t^-) = d_2) \frac{\Sigma_t(d_1,d_2)}{g^2_{0:t^-}}(\bar{A}(t^-), \bar{L}(t^-))\right],$$

where

$$\Sigma_t(d_1,d_2)(\bar{A}(t^-), \bar{L}(t^-)) = \mathbb{E}[\left(\bar{Q}_{t^+}^{d_1} - \bar{Q}_t^{d_1}\right)\left(\bar{Q}_{t^+}^{d_2} - \bar{Q}_t^{d_2}\right)|\bar{A}(t^-), \bar{L}(t^-)]$$

is the conditional covariance of $\bar{Q}_{t^+}^{d_1}$ and $\bar{Q}_{t^+}^{d_2}$, given $(\bar{A}(t^-), \bar{L}(t^-))$. Note that this can be obtained by regressing this cross-product on $(\bar{A}(t^-), \bar{L}(t^-))$. The latter sum can be further worked out giving us

$$\sigma_t^2 = \sum_{d_1\in\mathcal{D}} h_1(d_1,t) \mathbb{E} Z_{d_1}(d_1,t),$$

where

$$Z(d_1,t) = \left(\sum_{d_2\in\mathcal{D}} h_1(d_2,t)\mathbb{I}(d_1(\bar{L}(t^-)) = d_2(\bar{L}(t^-)))\right) \frac{\Sigma_t(d_1,d_2)}{g_{0:t^-}}\left(d_{1,t^-}(\bar{L}(t^-)), \bar{L}(t^-)\right),$$

so that the counterfactual of Z_t under intervention d_1 is given by:

$$Z_{d_1}(d_1,t) = \left(\sum_{d_2\in\mathcal{D}} h_1(d_2,t)\mathbb{I}\left(d_1\left(\bar{L}_{d_1}(t^-)\right) = d_2\left(\bar{L}_{d_1}(t^-)\right)\right)\right) \frac{\Sigma_t(d_1,d_2)}{g_{0:t^-}}\left(d_{1,t^-}\left(\bar{L}_{d_1}(t^-)\right), \bar{L}_{d_1}(t^-)\right).$$

With this expression, we can now use a TMLE to estimate $\mathbb{E} Z_{d_1}(d_1,t)$ for each $d_1 \in \mathcal{D}$ by using the longitudinal data structure with final outcome $Z(d_1,t)$, for each d_1 separately. To create the observed outcome $Z(d_1,t)$, we need a fit of the treatment mechanism $g_{A(m)} : m = 0, 1, \ldots, t^-$, evaluated at $\bar{A}(t^-) = d_{t^-}(\bar{L}(t^-))$, and for each rule compatible with d_1 (for that subject), we need to have an estimate of $\Sigma_t(d_1,d_2)$. Thus, given *a priori* estimates of the full treatment mechanism and all $(\Sigma_t(d_1,d_2) : d_1, d_2 \in \mathcal{D})$, we can construct this observed outcome $Z(d_1,t)$ and run the TMLE.

A.3 Estimation of the variance of the EIF

The aforementioned approach defines for each time point t and each rule d an observed longitudinal outcome $Z(d, t)$, where $Z(d, t)$ is a function of $(\bar{A}(t), \bar{L}(t))$. The TMLE of $\mathbb{E}Z_d(d, t)$ can then be computed based on the longitudinal data structure $(L(0), A(0), \ldots, L(t), A(t), Z(d, t))$ for each d and each $t \in \{0, 1, \ldots, K + 1\}$. As a result, we have that $\sigma_t^2 = \sum_{d \in \mathcal{D}} h_1(d, t)\mathbb{E}Z_d(d, t)$ and

$$\begin{aligned}\sigma^2 &= \sum_{t=0}^{K+1} \sigma_t^2 \\ &= \sum_{d \in \mathcal{D}} \left(\sum_{t=0}^{K+1} h_1(d, t)\mathbb{E}Z_d(d, t) \right) \\ &= \sum_{d \in \mathcal{D}} \mathbb{E}\left[\sum_{t=0}^{K+1} h_1(d, t)Z_d(d, t) \right].\end{aligned}$$

Let us now define the counterfactual outcome

$$\bar{Z}_d(d) \equiv \sum_{t=0}^{K+1} h_1(d, t)Z_d(d, t),$$

and the corresponding observed outcome

$$\bar{Z}(d) \equiv \sum_{t=0}^{K+1} h_1(d, t)Z(d, t).$$

We could apply the TMLE to estimate $\mathbb{E}\bar{Z}_d(d)$ based on the longitudinal data structure $(L(0), A(0), \ldots, L(K), A(K), \bar{Z}(d, K + 1))$, for each $d \in \mathcal{D}$, and use that

$$\sigma^2 = \sum_{d \in \mathcal{D}} \mathbb{E}\bar{Z}_d(d).$$

In applying TMLE here, we should be using that

$$\mathbb{E}[\bar{Z}_d|\bar{A}(m), \bar{L}(m)] = \sum_{t \leq m} h_1(d, t)Z(d, t) + \mathbb{E}\left[\sum_{t > m} h_1(d, t)Z(d, t)|\bar{A}(m), \bar{L}(m) \right].$$

To start with, let

$$\bar{Q}_d^{Z(K+1)} = \mathbb{E}[\bar{Z}(d)|\bar{A}(K), \bar{L}(K)] = \sum_{t \leq K} h_1(d, t)Z(d, t) + \mathbb{E}[h_1(d, K + 1) + Z(d, K + 1)|\bar{A}(K), \bar{L}(K)].$$

Denote the last conditional expectation with $\bar{Q}_d^{Z(K+1),d}$ so that

$$\bar{Q}_d^{Z(K+1)} = \sum_{t \leq K} h_1(d, t)Z(d, t) + \bar{Q}_d^{Z(K+1),d}.$$

Then,

$$\begin{aligned}\bar{Q}_d^{Z(K)} &= \mathbb{E}[\bar{Q}_d^{Z(K+1)}|\bar{A}(K - 1), \bar{L}(K - 1)] \\ &= \sum_{t \leq K-1} h_1(d, t)Z(d, t) + \mathbb{E}[h_1(d, K)Z(d, K) + \bar{Q}_d^{Z(K+1),d}|\bar{A}(K - 1), \bar{L}(K - 1)].\end{aligned}$$

Again, denote the latter conditional expectation by $\bar{Q}_d^{Z(K),d}$ so that

$$\bar{Q}_d^{Z(K)} = \sum_{t \leq K-1} h_1(d, t)Z(d, t) + \bar{Q}_d^{Z(K),d}.$$

Then,

$$\bar{Q}_d^{Z(K-1)} = \mathbb{E}[\bar{Q}_d^{Z(K)}|\bar{A}(K-2), \bar{L}(K-2)]$$
$$= \sum_{t \le K-2} h_1(d,t)Z(d,t) + \mathbb{E}[h_1(d,K-1)Z(d,K-1) + \bar{Q}_d^{Z(K),d}|\bar{A}(K-2), \bar{L}(K-2)].$$

Again, denote the latter conditional expectation with $\bar{Q}_d^{Z(K-1),d}$ so that

$$\bar{Q}_d^{Z(K-1)} = \sum_{t \le K-2} h_1(d,t)Z(d,t) + \bar{Q}_d^{Z(K-1),d}.$$

This is then iterated:

$$\bar{Q}_d^{Z(m)} = \sum_{t \le m-1} h_1(d,m)Z(d,m) + \bar{Q}_d^{Z(m),d},$$

where $\bar{Q}_d^{Z(m),d} = \mathbb{E}[h_1(d,m)Z(d,m) + \bar{Q}_d^{Z(m+1),d}|\bar{A}(m-1), \bar{L}(m-1)]$.

Before we go to the next conditional expectation, we need to target with a parametric submodel, such as

$$\text{Logit } \bar{Q}_d^m(\varepsilon) = \text{Logit } \bar{Q}_d^m + \varepsilon \frac{\mathbb{I}(\bar{A}(m-1) = d(\bar{L}(m-1)))}{g_{0:m-1}}.$$

In this way, we will only have to run one TMLE for each rule d, which still utilizes that the outcome is a sum of outcomes that are known for histories including that outcome.

B TMLE bootstrap consistency

Theorem 1. *Let P_n^0 be the initial estimator and $P_n^0(\varepsilon_n)$ be the parametric TMLE update so that $P_n D^*(P_n^0(\varepsilon_n)) = 0$. Suppose the following*:

(B0) *$D^*(P_n^0)$ falls in a P_0 Donsker class with probability tending to* 1,

(B1) *$R_0(P_n^0(\varepsilon_n), P_0) = o_P(n^{-1/2})$,*

(B2) *$P_0(D^*(P_n^0(\varepsilon_n)) - D^*(P_0))^2 \xrightarrow{p} 0$.*

Then,

$$n^{1/2}(\Psi(P_n^0(\varepsilon_n)) - \Psi(P_0)) = n^{1/2}P_n D^*(P_0) + o_P(1) \to N(0, \sigma_0^2 = P_0 D^{*2}(P_0)),$$

where the empirical measure of the bootstrap sample $O_i^{\#} \overset{iid}{\sim} P_n$ is denoted with $P_n^{\#}$ and $\varepsilon_n^{\#}$ is the maximum likelihood estimate of ε for $P_n^0(\varepsilon)$ based on $P_n^{\#}$.

Moreover, let now $P_n^0(\varepsilon_n^{\#})$ be TMLE update based on bootstrap sample $P_n^{\#}$ so that $P_n^{\#} D^*(P_n^0(\varepsilon_n^{\#})) = 0$.

We have $n^{1/2}\Psi(P_n^0(\varepsilon_n^{\#})) - \Psi(P_n^0(\varepsilon_n)) = n^{1/2}(P_n^{\#} - P_n)D^*(P_0) + o_P(1)$, which (conditional on P_n) converges to $N(0, \sigma_0^2)$. This proves the consistency of the non-parametric bootstrap for the TMLE that treats the initial estimator P_n^0 as fixed.

Note that we do not require more than B0, B1, and B2, which are the assumptions under which TMLE is asymptotically linear and efficient.

Proof. Let P_n^0 be the initial estimator. Let $P_n^* = P_n^0(\varepsilon_n)$ be the TMLE and $\Psi(P_n^0(\varepsilon_n))$ be the plug in for the TMLE. Let us first analyze the TMLE.

By the usual identity for TMLE, we have

$$\Psi(P_n^0(\varepsilon_n)) - \Psi(P_0) = (P_n - P_0)D^*(P_n^0(\varepsilon_n)) + R_0(P_n^0(\varepsilon_n), P_0).$$

Since for TMLE we assume the initial estimator converges fast enough, we have $R_0(P_n^0(\varepsilon_n), P_0) = o_P(n^{-1/2})$ (B1). Moreover,

$$(P_n - P_0)D^*(P_n^0(\varepsilon_n)) = (P_n - P_0)(D^*(P_n^0(\varepsilon_n)) - D^*(P_0)) + (P_n - P_0)D^*(P_0).$$

Our assumption of Donsker classes (B0) and (B1) results in the first term $(P_n - P_0)D^*(P_n^0(\varepsilon_n))$ being an empirical process term $o_P(n^{-1/2})$. Thus, we now have

$$\Psi(P_n^0(\varepsilon_n)) - \Psi(P_0) = (P_n - P_0)D^*(P_0) + o_P(n^{-1/2})$$

proving asymptotic efficiency. The remainder of the proof essentially just repeats this proof for the TMLE that treats P_n^0 as fixed but now we base this TMLE on a sample from P_n, whose empirical distribution we denote with $P_n^\#$ and the target is $\Psi(P_n^0(\varepsilon_n))$ (i.e., the truth under sampling from P_n). This results in the TMLE on the bootstrap sample being $\Psi(P_n^0(\varepsilon_n^\#))$ and

$$\Psi(P_n^0(\varepsilon_n^\#)) - \Psi(P_0) = (P_n^\# - P_0)D^*(P_n^0(\varepsilon_n^\#)) + R_{P_0}(P_n^0(\varepsilon_n^\#), P_0)$$
$$\Psi(P_n^0(\varepsilon_n)) - \Psi(P_0) = (P_n - P_0)D^*(P_n^0(\varepsilon_n)) + R_{P_0}(P_n^0(\varepsilon_n), P_0).$$

Subtracting the two equations gives

$$\Psi(P_n^0(\varepsilon_n^\#)) - \Psi(P_n^0(\varepsilon_n)) = (P_n^\# - P_0)D^*(P_n^0(\varepsilon_n^\#)) - (P_n - P_0)D^*(P_n^0(\varepsilon_n)) + o_P(n^{-1/2})$$

since both remainder terms are $o_P(n^{-1/2})$. Note that the left-hand side is our bootstrapped centered estimator. However, since $(P_n^\# - P_0)D^* = (P_n^\# - P_n)D^* + (P_n - P_0)D^*$, the difference of the two empirical process terms equals

$$(P_n^\# - P_n)D^*(P_n^0(\varepsilon_n^\#)) + (P_n - P_0)D^*(P_n^0(\varepsilon_n^\#)) - D^*(P_n^0(\varepsilon_n)).$$

The second term is an empirical process term $o_P(n^{-1/2})$ due to $\varepsilon_n^\# - \varepsilon_n$ converging to zero and Donsker class condition (B0, B1). Thus, we have

$$\Psi(P_n^0(\varepsilon_n^\#)) - \Psi(P_n^0(\varepsilon_n)) = (P_n^\# - P_n)D^*(P_n^0(\varepsilon_n^\#)) + o_P(n^{-1/2}).$$

Finally, the latter empirical process term is

$$(P_n^\# - P_n)D^*(P_0) + (P_n^\# - P_n)(D^*(P_n^0(\varepsilon_n^\#)) - D^*(P_0)),$$

and the latter term is again $o_P(n^{-1/2})$ by the consistency of $P_n^0(\varepsilon_n^\#)$ to P_0 and Donsker holding automatically since P_n^0 is fixed to only a parametric model. We have therefore shown

$$\Psi(P_n^0(\varepsilon_n^\#)) - \Psi(P_n^0(\varepsilon_n)) = (P_n^\# - P_n)D^*(P_0) + o_P(n^{-1/2})$$

since P_n^0 is fixed, so that the class of functions covering $D^*(P_n^0(\varepsilon_n^\#))$, conditional on P_n, is just a parametric class. Conditional on P_n and multiplying by $n^{1/2}$, we have that this converges to the same $N(0, \sigma^2)$ proving consistency for the bootstrap. □