Variable importance measures for heterogeneous causal effects

Oliver Hines[1], Karla Diaz-Ordaz[1], and Stijn Vansteelandt[1,2]

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, U.K.
[2]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

February 22, 2022

Abstract

The conditional average treatment effect (CATE) of a binary exposure on an outcome is often studied on the basis that personalised treatment decisions lead to better clinical outcomes. The CATE, however, may be a complicated function of several covariates representing patient characteristics. As such, clinicians rely on researchers to summarise insights related to treatment effect heterogeneity. Clinical research usually focuses on the ATE, which averages the CATE across all patient characteristics, or considers the treatment effect in patient strata, though strata are rarely chosen systematically. Here we present new nonparametric treatment effect variable importance measures (TE-VIMs). These may guide clinicians as to which patient characteristics are important to consider when making treatment decisions, and help researchers stratify patient populations for further study. TE-VIMs generalise recent regression-VIMs, viewed as nonparametric analogues to ANOVA statistics. Moreover, TE-VIMs are not tied to a particular model, thus are amenable to data-adaptive (machine learning) estimation of the CATE, itself an active area of research. Estimators for the proposed statistics are derived from their efficient influence curves and these are illustrated through a simulation study and an applied example.

1 Introduction

Heterogeneous causal effects are important in medical and social sciences and play an important role in informing policy and optimizing treatment decisions. Heterogeneity is commonly studied through the conditional average treatment effect (CATE) $\tau(x) \equiv E(Y^1 - Y^0|X = x)$, where Y^a denotes the outcome that would be observed if exposure had taken the value $A = a$, and X is a vector of covariates.

CATE estimation is difficult, especially when components of X are continuous and $\tau(x)$ represents an infinite dimensional parameter (i.e. a function). This problem has received significant attention in recent years with an emphasis on flexible machine learning methods (Abrevaya et al., 2015; Athey and Imbens, 2016; Nie and Wager, 2017; Kallus et al., 2018; Wager and Athey, 2018; Künzel et al., 2019; Kennedy, 2020). Much early work in causal inference instead focused on the average treatment effect (ATE) $E\{\tau(X)\}$, and weighted variations, as the target estimand. We view ATEs as scalar summaries of the CATE, which help answer scientific questions of interest. ATE estimation is simpler than CATE estimation since, under standard causal assumptions, efficient estimators can be constructed using the efficient influence curve under the nonparametric model, e.g. double machine learning (Chernozhukov et al., 2018) and targeted learning (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). Moreover, these estimators are not tied to a particular model, thus are amenable to data adaptive modelling of $\mu(a, x) \equiv E(Y|A = a, X = x)$, and the propensity score, $\pi(x) \equiv E(A|X = x)$.

In policy and decision making environments it is unfeasible to appeal to the CATE directly thus researchers are required to summarise insights related to treatment effect heterogeneity in both experimental and observational

data settings. One approach for communicating heterogeneity is to compare the estimated treatment effect in different population subgroups, see e.g. Rothwell (2005) for a review of subgroup analysis approaches in clinical trials. Such analyses also identify subgroups for potential future study which benefit most/least from exposure. Faced with many pre-exposure covariates, however, it is unclear how stratification variables should be selected and these are rarely chosen systematically.

Several existing CATE estimation strategies permit the analysis of variable importance. For instance, Wager and Athey (2018) extend the widely used 'random forest' algorithm (Breiman, 2001) to CATE estimation, which permits random forest variable importance measures to be applied to the CATE (Grömping, 2009). The resulting importance measures, however, are inherently tied to the modelling strategy and do not generalise. Instead, we propose treatment effect variable importance measures (TE-VIMs), which are nonparametrically defined scalar summary statistics intended to measure the importance of subsets of covariates in predicting the CATE. Our proposed estimands are based on,

$$\Theta_s \equiv E[\text{var}\{\tau(X)|X_{-s}\}] = \text{var}\{\tau(X)\} - \text{var}\{\tau_s(X)\} \tag{1}$$

where $X \in \mathbb{R}^p$ represents a p-dimensional covariate vector, the symbol u_{-s} denotes the vector of all the components of u with index not in $s \subseteq \{1, ..., p\}$, and $\tau_s(x) \equiv E\{\tau(X)|X_{-s} = x_{-s}\}$ denotes the CATE conditional on X_{-s}. We interpret $\Theta_s \geq 0$ as the additional treatment effect heterogeneity explained by conditioning on X_s as well as conditioning on X_{-s}, where for a vector u, we denote by u_s the vector of all components of u with index in s.

The estimand $\text{var}\{\tau(X)\}$ represents the variance of treatment effect (VTE) (Levy et al., 2021), a recently proposed measure of treatment effect heterogeneity, which captures the extent to which varying treatment outcomes can be explained by observed covariates. The estimand Θ_s is therefore understood as a difference in VTEs, quantifying the amount by which the VTE changes when variables in the set s are excluded from the model. The proposed TE-VIMs rescale Θ_s by the VTE to express this difference as a proportion,

$$\Psi_s \equiv \frac{\Theta_s}{\text{var}\{\tau(X)\}} = 1 - \frac{\text{var}\{\tau_s(X)\}}{\text{var}\{\tau(X)\}}. \tag{2}$$

Assuming that the VTE is non-zero, we interpret $\Psi_s \in [0, 1]$ as the proportion of treatment effect heterogeneity explained by X_{-s} compared with X. This is interpretation is analogous to the familiar coefficient of determination (R^2 statistic).

We recommend that TE-VIM inference could be incorporated into a more broad treatment effect analysis, where primary interest is in inferring the ATE and VTE. We believe that VTE inference should form part of a primary analysis, since it is possible that the population ATE is zero, but some (or indeed all) individuals experience a large CATE. One may then infer TE-VIMs as part of a secondary analysis, once treatment effect heterogeneity has been established, since TE-VIM estimands are not of scientific interest when there is little variability in the CATE to account for. In particular, under treatment effect homogeneity, $\tau(x)$ is constant, $\Theta_s = 0$ for all covariate subsets and Ψ_s breaks down because $\text{var}\{\tau(X)\} = 0$.

The proposed TE-VIMs connect VTE estimands (Levy et al., 2021) to recently proposed regression-VIMs (Williamson et al., 2021a,b; Zhang and Janson, 2020), also referred to as 'leave out covariates' (Verdinelli and Wasserman, 2021; Lei et al., 2018). In the regression setting, interest is in quantifying the importance of sets of variables in predicting the conditional mean outcome $\mu(x) \equiv E(Y|X = x)$, rather than the CATE $\tau(x)$. Williamson et al. (2021a) propose regression VIMs based on $\theta_s = E[\text{var}\{\mu(X)|X_{-s}\}]$, which is analogous to Θ_s in the current work. In this way, our work represents a step towards extending regression-VIMs to the analysis of more general statistical functionals.

In Section 2 we motivate TE-VIMs, and provide estimators which are efficient under the nonparametric model. These rely on estimating working models, relating our proposal to the DR-learner of the CATE through an interpreation based on so-called pseudo-outcomes (Kennedy, 2020; Luedtke and van der Laan, 2016; van der Laan, 2013). Experimental results on simulated data are provided in Section 3 and Section 4 demonstrates an application to the AIDS Clinical Trials Group Protocol 175 (Hammer et al., 1996).

2 Methodology

2.1 Motivating the estimand

Suppose we have n iid observations, $(z_1, ..., z_n)$ of a random variable Z distributed according to an unknown distribution P, such that Z consists of (Y, A, X), where $Y \in \mathbb{R}$ is an 'outcome', $A \in \{0, 1\}$ is an 'exposure' and $X \in \mathbb{R}^p$ is a p-dimensional vector of covariates. The CATE is defined as $\tau(x) \equiv E(Y^1 - Y^0|X = x)$, where Y^a denotes the outcome that would be observed if exposure had taken the value a, (Rubin, 1974). Under standard identification assumptions of consistency ($A = a \implies Y = Y^a$), conditional exchangeability ($Y^a \perp\!\!\!\perp A|X$), and positivity ($0 < \pi(X) < 1$), the CATE is identified by $\tau(x) = \mu(1, x) - \mu(0, x)$.

Assume that $||\tau|| < \infty$, where $||f|| \equiv E\{f(X)^2\}^{1/2}$ is the $L_2(P)$ norm. With this choice our estimand is finite and well defined, since

$$\Theta_s = ||\tau - \tau_s||^2 = E[\{\tau(X) - \tau_s(X)\}^2] < \infty,$$

Notice that the VTE is Θ_p, where, in a slight abuse of notation, p denotes the index set $\{1, ..., p\}$ and τ_p is the ATE. We further assume that the VTE is non-zero, i.e. $\Theta_p > 0$ and since $\Theta_p \geq \Theta_s$, it follows that $\Psi_s = \Theta_s/\Theta_p \in [0, 1]$.

The regression-VIM in Williamson et al. (2021a) is analogous to our proposal, in the sense that the former is recovered when $\tau(x)$ is replaced with $\mu(x)$ and $\tau_s(x)$ is replaced with $\mu_s(x) \equiv E(Y|X_{-s} = x_{-s})$. In other words, when the observed outcome Y plays the role of the causal contrast, $Y^1 - Y^0$. Specifically, they consider,

$$\theta_s \equiv E[\text{var}\{\mu(X)|X_{-s}\}] = E[\{\mu(X) - \mu_s(X)\}^2]$$

which is analogous to Θ_s. The two proposals, however, differ in how this mean conditional variance is scaled. Williamson et al. (2021a) consider scaling by the outcome variance, i.e. by defining $\psi_s \equiv \theta_s/\text{var}(Y)$, whereas we scale by the VTE, i.e. $\Psi_s = \Theta_s/\Theta_p$, hence our estimand is proportional to the more direct analogue, $\Psi_s \propto \Theta_s/\text{var}(Y^1 - Y^0)$, with a proportionality constant of $\Theta_p/\text{var}(Y^1 - Y^0)$ for all covariate sets. We scale by the VTE because the treatment effect variance, $\text{var}(Y^1 - Y^0)$, is generally not identifiable without strong assumptions (Levy et al., 2021; Ding et al., 2016; Heckman et al., 1997). The VTE, however, is a convenient scaling parameter which bounds $\Psi_s \in [0, 1]$, aiding interpretability since, when the VTE is non-zero, Ψ_s behaves like a coefficient of determination (R^2).

In practice, the scale factor makes little difference to the interpretation of our estimands, since investigators are likely to compare the relative importance of covariate sets s and s' by comparing the magnitudes of Ψ_s and $\Psi_{s'}$. This approach is demonstrated in Section 4, where the importance of each covariate is ranked. Alternatively, one might compare the importance of sets of related covariates e.g. biological factors vs. non-biological factors.

2.2 Efficient estimation

Next we consider the efficient influence curves (ICs) of Θ_s and Ψ_s under the nonparametric model. Briefly, ICs are model-free, mean zero, functionals that characterise the sensitivity of an estimand to small changes in the data generating law. As such, ICs are useful for constructing efficient estimators and determining their asymptotic distribution, see e.g. Hines et al. (2022); Fisher and Kennedy (2020) for an introduction to these methods. Letting $z = (y, a, x)$ denote a single observation of Z, we first consider the IC and efficient estimators of Θ_s. In the Appendix we derive that the IC of Θ_s is,

$$\phi_s(z) = \{\varphi(z) - \tau_s(x)\}^2 - \{\varphi(z) - \tau(x)\}^2 - \Theta_s \tag{3}$$

where,

$$\varphi(z) \equiv \{y - \mu(a,x)\}\frac{a - \pi(x)}{\pi(x)\{1-\pi(x)\}} + \mu(1,x) - \mu(0,x).$$

This is the 'pseudo-outcome' of the DR-learner (Kennedy, 2020), also called the augmented inverse propensity weighted score (Robins et al., 1994), with the interpretation that $\varphi(Z)$ acts like the causal contrast, $Y^1 - Y^0$, since $\tau_s(x) = E\{\varphi(Z)|X_{-s} = x_{-s}\}$. This interpretation holds also in the present context. To see why, we compare the IC in (3) with that of θ_s given by Williamson et al. (2021a),

$$\{y - \mu_s(x)\}^2 - \{y - \mu(x)\}^2 - \theta_s$$

The IC of θ_s has the same form as (3), since the latter is recovered by replacing the outcome y with the pseudo-outcome, $\varphi(z)$, replacing the conditional mean outcomes with CATEs, i.e. conditional means of $\varphi(Z)$, and replacing θ_s with Θ_s, hence the pseudo-outcome plays the role of the unobserved causal contrast, $Y^1 - Y^0$.

Efficient estimating equations estimators can be derived from ICs by setting (an estimate of) the sample mean IC to zero. In the current setting, this strategy is equivalent to the so-called one-step correction which we outline in the Appendix. For Θ_s and θ_s, we obtain the estimators

$$\begin{aligned} \hat{\Theta}_s &\equiv n^{-1}\sum_{i=1}^{n}\{\hat{\varphi}(z_i) - \hat{\tau}_s(x_i)\}^2 - \{\hat{\varphi}(z_i) - \hat{\tau}(x_i)\}^2 \\ \hat{\theta}_s &\equiv n^{-1}\sum_{i=1}^{n}\{y_i - \hat{\mu}_s(x_i)\}^2 - \{y_i - \hat{\mu}(x_i)\}^2 \end{aligned} \tag{4}$$

where $\hat{\tau}(z)$ and $\hat{\tau}_s(z)$ are consistent CATE estimators and we define the pseudo-outcome plug-in estimator,

$$\hat{\varphi}(z) \equiv \{y - \hat{\mu}(a,x)\}\frac{a - \hat{\pi}(x)}{\hat{\pi}(x)\{1-\hat{\pi}(x)\}} + \hat{\mu}(1,x) - \hat{\mu}(0,x)$$

which is based on estimates of $\mu(a,x)$ and $\pi(x)$ obtained from an independent sample. In practice, a cross-fitting approach (see Section 2.3) may be used to obtain the fitted models and evaluate $\hat{\Theta}_s$ from a single sample (Zheng and van der Laan, 2011; Chernozhukov et al., 2018). The estimator $\hat{\Theta}_s$ is a new proposal of this work and is analogous to $\hat{\theta}_s$ which was proposed by Williamson et al. (2021a).

For the index set $s = p$, then $\hat{\Theta}_p$ is an estimator of the VTE, which is distinct from the targeted maximum likelihood estimation (TMLE) VTE estimator proposed by (Levy et al., 2021). Both are based on initial estimates of $\mu(a,x)$ and $\pi(x)$ and are regular asymptotically linear in the sense of removing plug-in bias by ensuring that the sample mean of the estimated IC is zero. The TMLE estimator achieves this by replacing the initial estimates $\hat{\mu}(a,x)$, with 'retargeted' estimates $\hat{\mu}^*(a,x)$. These retargeted estimates are used to estimate the CATE as $\hat{\tau}^*(x) \equiv \hat{\mu}^*(1,x) - \hat{\mu}^*(0,x)$, and ATE as the sample mean of $\hat{\tau}^*(x)$. This strategy ensures that the TMLE VTE estimator is a plug in estimator, see e.g. Hines et al. (2022) for an introduction.

Conversely, our VTE estimator follows an estimating equations/ one-step bias reduction strategy and is unconventional by not requiring that the CATE estimates and pseudo-outcome estimates are obtained from the same plug-in models. The advantages of doing so can be seen by considering the learning rate assumptions which we use to bound the second order remainder term of our Θ_s estimator:

(A1) The differences $\tau(x) - \hat{\tau}(x)$, $\tau_s(x) - \hat{\tau}_s(x)$, and $\tau_p - \hat{\tau}_p$ are all $o_P(n^{-1/4})$ in $L_2(P)$ norm.

(A2) The propensity score and outcome estimators are 'double robust' in the sense that $\{\pi(x) - \hat{\pi}(x)\}\{\mu(a,x) - \hat{\mu}(a,x)\}$ is $o_P(n^{-1/2})$ in $L_2(P)$ norm.

A similar assumption to (A1) appears in Williamson et al. (2021a) related to the outcome models $\hat{\mu}(x)$ and $\hat{\mu}_s(x)$, in the regression-VIM setting where the outcome is observed. In our setting, the pseudo-outcomes are

not observed and we require (A2) to control the error in estimating these pseudo-outcomes. (A2) is the standard 'double robust' assumption, which appears when considering efficient estimators of the ATE.

Supposing that we use the plug-in CATE estimates $\hat{\tau}(x) = \hat{\mu}(1,x) - \hat{\mu}(0,x)$ then (A1) is satisfied provided that $\hat{\mu}(a,x) - \mu(a,x)$ is $o_P(n^{-\delta})$ in $L_2(P)$ norm, with $\delta > 1/4$. (A2) then implies that $\pi(x) - \hat{\pi}(x)$ must be at least $o_P(n^{-1/2+\delta})$ in $L_2(P)$ norm. In other words, our propensity score model can converge at a slower rate, provided the outcome model converges at a faster rate, but the converse is not true. This is unsatisfying for example in clinical trial settings, where the exposure is randomised and the propensity score model is known, since the plug-in CATE estimator would still require $n^{1/4}$ rate convergence of the outcome model.

By separating the CATE and pseudo-outcome modelling assumptions, however, one is free to use CATE estimates which exploit propensity score modelling e.g. the double robust learner (DR-learner) of Kennedy (2020). In particular, this means that double robust VTE estimates could be obtained, marking an improvement over the TMLE VTE estimator, which is not double robust. The cost of this robustness, however, is that our VTE estimator is not itself a plug-in estimator, and therefore does not necessarily respect the parameter space, while the TMLE estimator ensures that $\hat{\Theta}_p^{(TMLE)} \geq 0$.

Our main goal is to consider efficient estimation of the TE-VIM Ψ_s, which has IC,

$$\Phi_s(z) = \{\phi_s(z) - \Psi_s \phi_p(z)\}/\Theta_p \tag{5}$$

where $\phi_p(.)$ denotes (3) for the index set $s = p$. This IC implies an estimating equations estimator, $\hat{\Psi}_s = \hat{\Theta}_s/\hat{\Theta}_p$, provided that $\Theta_p \neq 0$.

Since the estimators, $\hat{\Theta}_s$ and $\hat{\theta}_s$ are analogous, we expect that they share similar properties. One such property concerns the behaviour of the estimator under the zero-importance null hypothesis, $H_0 : \Theta_s = 0$, i.e. when $\tau(x) = \tau_s(x)$. In practice, this hypothesis corresponds to treatment effect homogeneity over X_s given X_{-s}, thus the ICs in (3) and (5) are exactly zero. The fact that the IC in (3) degenerates in this way makes the hypothesis, $H_0 : \Theta_s = 0$ difficult to test, since Wald based tests, i.e. those based on the asymptotic distribution of $\hat{\Theta}_s$ will generally be conservative. Usually the IC characterises the asymptotic distribution of the efficient estimator, as in Theorem 1 below. Under H_0, however, one must consider higher-order pathwise derivatives of the estimand see e.g. Carone et al. (2018). This remains generally an open problem and, for this reason, Theorem 1 only considers the behaviour of the estimator when $\Psi_s \in (0, 1]$.

Theorem 1 *Assume that $\Theta_p \neq 0$. Under (A1), (A2) and regularity assumptions given in the Appendix, $\hat{\Psi}_s$ is asymptotically linear with IC, $\Phi_s(Z)$, and hence $\hat{\Psi}_s$ converges to Ψ_s in probability, and for $\Psi_s \in (0, 1]$ then $n^{1/2}(\hat{\Psi}_s - \Psi_s)$ converges in distribution to a mean-zero normal random variable with variance $E\{\Phi_s(Z)^2\}$.*

2.3 Proposed learning algorithms

The estimator $\hat{\Theta}_s$ is indexed by the choice of pseudo-outcome and CATE estimators. Generally, we are not constrained to any particular learning method. Throughout, we consider a plug-in estimator for $\varphi(.)$, where models for $\mu(a,x)$ and $\pi(x)$ are obtained from the same sample, and these are used to construct $\hat{\varphi}(z)$.

This strategy is used by the ATE augmented inverse propensity weighted estimator, and the DR-learner (Kennedy, 2020; Luedtke and van der Laan, 2016; van der Laan, 2013), however the latter usually requires that $\hat{\mu}(.,.)$ and $\hat{\pi}(.)$ are obtained from two independent samples. We consider two approaches to learning $\tau(.)$. The first is to use the plug-in estimator $\hat{\tau}(x) = \hat{\mu}(1,x) - \hat{\mu}(0,x)$, referred to as T-learner (Künzel et al., 2019).

The second relies on the observation that, once an independent pseudo-outcome learner, $\hat{\varphi}(.)$ has been established, an estimator for $\tau(.)$ and $\tau_s(.)$ can be constructed from a regression of $\hat{\varphi}(Z)$ on X and X_{-s} respectively. This is the principle behind the DR-learner (Kennedy, 2020), which uses a double cross fitting strategy for nuisance model estimation (Newey and Robins, 2018). We consider this approach for estimating $\tau(.)$, in the setting where $\hat{\varphi}(.)$ is fitted on the same sample that is used for the subsequent regression, i.e. without cross fitting.

We expect the DR-learning approach to outperform the T-learning approach in finite samples due to its robustness properties and since the DR-learner explicitly seeks to minimise the term, $n^{-1}\sum_{i=1}^{n}\{\hat{\varphi}(z_i)-\hat{\tau}(x_i)\}^2$, which appears in (4). This term can be problematic in practice, since it may give negative $\hat{\Theta}_s$ estimates, when $\hat{\tau}(.)$ converges slowly to $\tau(.)$. Asymptotically, however, we cannot expect a difference unless a benefit can be shown, e.g. under outcome model misspecification.

Moreover, given an independent estimator $\hat{\varphi}(.)$, one is effectively in a setting where pseudo-outcomes are observed, and our estimators are analogous to those in Williamson et al. (2021a). In that work, the authors found that regressing the observed outcome on X and X_{-s} returns conditional mean models which do not take into account that the two conditional means are related, generally resulting in incompatible estimates. Their solution was to first regress the outcome on X then regress predictions from the resulting conditional mean model on X_{-s}. We therefore propose a similar approach to learning $\hat{\tau}_s(.)$.

The proposed working function estimators are implemented in Algorithms 1 and 2 below. These algorithms return pseudo-outcome and CATE estimates, $\{\hat{\varphi}_i\}_{i=1}^n, \{\hat{\tau}_i\}_{i=1}^n,$, $\{\hat{\tau}_{s,i}\}_{i=1}^n$, and $\{\hat{\tau}_{p,i}\}_{i=1}^n$, which can be used to obtain $\hat{\Psi}_s = \hat{\Theta}_s/\hat{\Theta}_p$, with variance estimated by $n^{-2}\sum_{i=1}^{n}\hat{\phi}_i^2$, where,

$$\hat{\Theta}_s = n^{-1}\sum_{i=1}^{n}\{\hat{\varphi}_i-\hat{\tau}_{s,i}\}^2-\{\hat{\varphi}_i-\hat{\tau}_i\}^2$$
$$\hat{\phi}_i = \frac{1}{\hat{\Theta}_p}\left[\{\hat{\varphi}_i-\hat{\tau}_{s,i}\}^2-\hat{\Psi}_s\{\hat{\varphi}_i-\hat{\tau}_{p,i}\}^2+(\hat{\Psi}_s-1)\{\hat{\varphi}_i-\hat{\tau}_i\}^2\right]$$

and $\hat{\Theta}_p$ is defined in a similar manner. Algorithm 2 uses a cross fitting regime to ensure that $\hat{\varphi}_i, \hat{\tau}_i, \hat{\tau}_{s,i}$, and $\hat{\tau}_{p,i}$ are constructed from working models which are not fitted using the ith observation. This is useful in controlling the so-called empirical process term, see e.g. Newey and Robins (2018); Hines et al. (2022). Both algorithms are also indexed by the choice of CATE learner in steps 2 and 3 of each algorithm respectively, with the substeps marked (A) and (B) referring to the T, and DR-learners. We note that where the algorithms require models to be 'fitted', any suitable regression method can be used.

Algorithm 1 - Without sample splitting

(1) Fit $\hat{\mu}(.,.)$ and $\hat{\pi}(.)$. Use these fitted models to obtain $\hat{\varphi}_i \equiv \hat{\varphi}(z_i)$.

(2) (A) Use the model for $\hat{\mu}(.,.)$ from Step 1, to obtain $\hat{\tau}(x) \equiv \hat{\mu}(1,x)-\hat{\mu}(0,x)$. Or (B) Fit $\hat{\tau}(.)$ by regressing $\hat{\varphi}(Z)$ on X. After doing (A) or (B), use the fitted models to obtain $\hat{\tau}_i \equiv \hat{\tau}(x_i)$.

(3) Fit $\hat{\tau}_s(.)$ by regressing $\hat{\tau}(X)$ on X_{-s}. Use the fitted model to obtain $\hat{\tau}_{s,i} \equiv \hat{\tau}_s(x_i)$.

(4) Repeat Step 3 for the covariate set p and (optionally) any other covariate sets of interest.

Algorithm 2 - With sample splitting

(1) Split the data into K folds.

(2) **For** each fold k: Fit $\hat{\mu}(.,.)$ and $\hat{\pi}(.)$ using the data set excluding fold k. Use these fitted models to obtain $\hat{\varphi}_i \equiv \hat{\varphi}(z_i)$ for i in fold k.

(3) (A) Use the model for $\hat{\mu}(.,.)$ from Step 2, to obtain $\hat{\tau}(x) \equiv \hat{\mu}(1,x)-\hat{\mu}(0,x)$. Or (B) Fit $\hat{\tau}(.)$ by regressing $\hat{\varphi}(Z)$ on X using the data excluding fold k. After doing (A) or (B), use the fitted models to obtain $\hat{\tau}_i \equiv \hat{\tau}(x_i)$ for i in fold k

(4) Fit $\hat{\tau}_s(.)$ by regressing $\hat{\tau}(X)$ on X_{-s} using the data excluding fold k. Use the fitted model to obtain $\hat{\tau}_{s,i} \equiv \hat{\tau}_s(x_i)$ for i in fold k.

(5) Repeat Step 4 for the covariate set p and (optionally) any other covariate sets of interest. **End for.**

3 Simulation study

In our simulation study we compared Algorithms 1A,1B, 2A and 2B on generated data in finite samples, using $K = 5$ fold sample splitting. We generated 1000 datasets of size $n \in \{500, 1000, 2000, 3000, 4000\}$ from the following structural equation model

$$\begin{aligned} X_1, X_2 &\sim \text{Uniform}(-1, 1) \\ A &\sim \text{Bernoulli}\{\text{expit}(-0.4X_1 + 0.1X_1X_2)\} \\ Y &\sim \mathcal{N}\left(\{X_1X_2 + 2X_2^2 - X_1\} + A\tau(X), 1\right) \end{aligned}$$

where the CATE is given by $\tau(X) = X_1^2(X_1 + 7/5) + 25X_2^2/9$. And the TE-VIMs take the true values $\Psi_1 \approx 0.32$ and $\Psi_2 \approx 0.68$.

For each dataset, $\hat{\Psi}_s$ was estimated along with its variance and associated Wald based (95%) confidence intervals for the index sets, $s = \{1\}, \{2\}$. Two regression model approaches were considered, the first used generalised additive models, as implemented through the `mgcv` package in R (Wood et al., 2016). These models use flexible spline smoothing with interaction terms and the propensity score model included a logistic link function. The second regression modelling approach used random forest learners available through the `ranger` package in R (Wright and Ziegler, 2017).

Figure 1 shows empirical estimates of the bias and variance of $\hat{\Psi}_1$ scaled by $n^{1/2}$ and n respectively, as well as 95% Wald based confidence-interval coverage probabilities. Similar plots for $\hat{\Psi}_2$ are in the Appendix. Comparing Algorithms 1 and 2 (i.e. no sample splitting vs sample splitting), we notice a greater difference in the results when random forest learning is used, with sample splitting generally reducing bias, increasing variance and improving confidence interval coverage.

Additionally, the DR-learning approach (Algorithm B) outperforms the T-learning approach (Algorithm A), by making better use of the propensity score model to improve estimation of the CATE. On the basis of these results, we recommend Algorithm 2B for TE-VIM learning.

4 HIV data analysis

We demonstrate our estimators on data from the AIDS Clinical Trials Group Protocol 175 (ACTG175) (Hammer et al., 1996), which consisted of 2139 patients infected with HIV whose CD4 T-cell count was in the range 200 to 500 mm^{-3}. Patients were randomised to 4 treatment groups: (i) zidovudine (ZDV) monotherapy, (ii) ZDV+didanosine (ddI), (iii) ZDV+zalcitabine, and (iv) ddI monotherapy. We compare treatment groups (iv) and (ii) as in Lu et al. (2013); Cui et al. (2020). These two groups are represented with the binary indicator, $A = 0, 1$, with $n = 561$ and $n = 522$ patients in each group respectively.

Previous studies have used ACTG175 data to analyse the causal effect of A on a survival time endpoint, and the data is available through the `speff2trial` package in R. We consider CD4 count at 20±5 weeks as a continuous outcome, Y, and consider 12 baseline covariates, 5 continuous: age, weight, Karnofsky score, CD4 count, CD8 count; and 7 binary: sex, homosexual activity (y/n), race (white/non-white), symptomatic status (symptomatic/asymptomatic), history of intravenous drug use (y/n), hemophilia (y/n), and antiretroviral history (experienced/naive).

TE-VIMs for each covariates were estimated using all algorithms with $K = 15$ folds. Propensity score estimates were obtained as the mean of the treatment indicator in the training set. This model is correctly specified since treatment is randomised. Other fitted models were obtained using the Super Learner (van der Laan et al., 2007), an ensemble learning method, implemented in the `SuperLearner` package in R. This used 15 cross validation folds, and a 'learner library' containing various routines (`mean`, `glm`, `glmnet`, `gam`, `xgboost`, `ranger`, `nnet`).

The cross-fitted augmented inverse propensity weighted estimate of the ATE was $29.1mm^{-3}$ (CI: 14.6, 43.7;

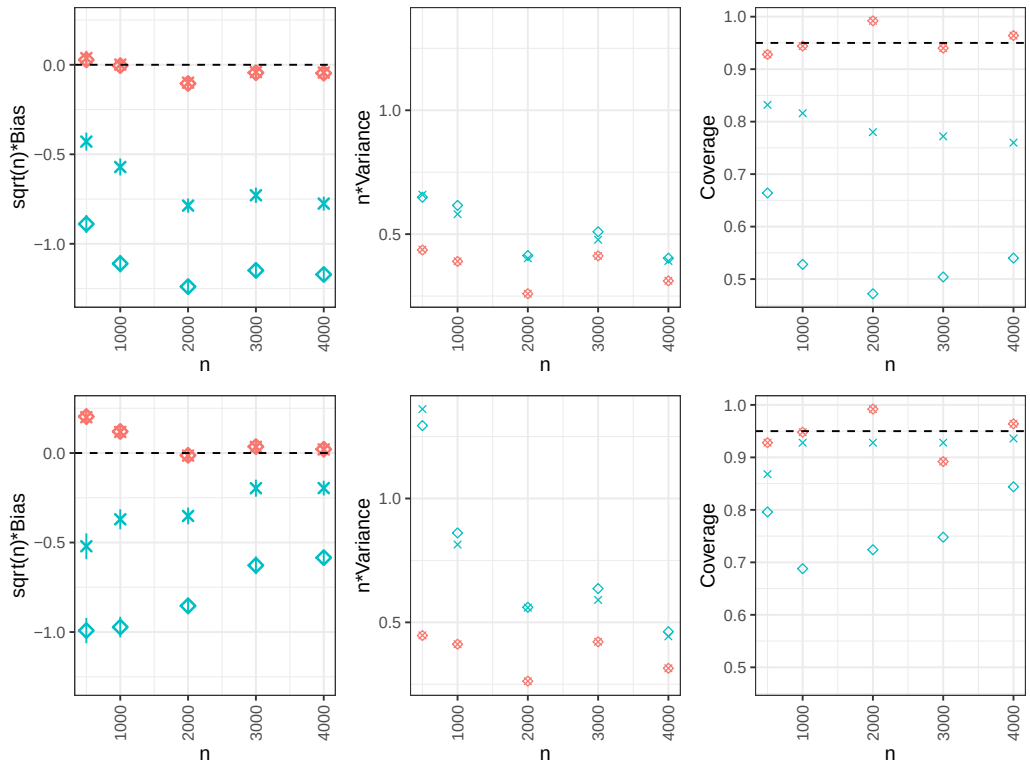

Figure 1: Bias, variance and coverage for $\hat{\Psi}_1$ using 1000 sampled datasets. Red and blue points indicate that working models are fitted using generalised additive modelling and random forests respectively. Top row of plots corresponds to Algorithm 1 (no sample splitting) and the bottom row corresponds to Algorithm 2 (sample splitting). Square and crossed points indicate that the algorithm used the T-learner and DR-learner respectively for CATE estimation. Similar plots for $\hat{\Psi}_2$ are given in the Appendix.

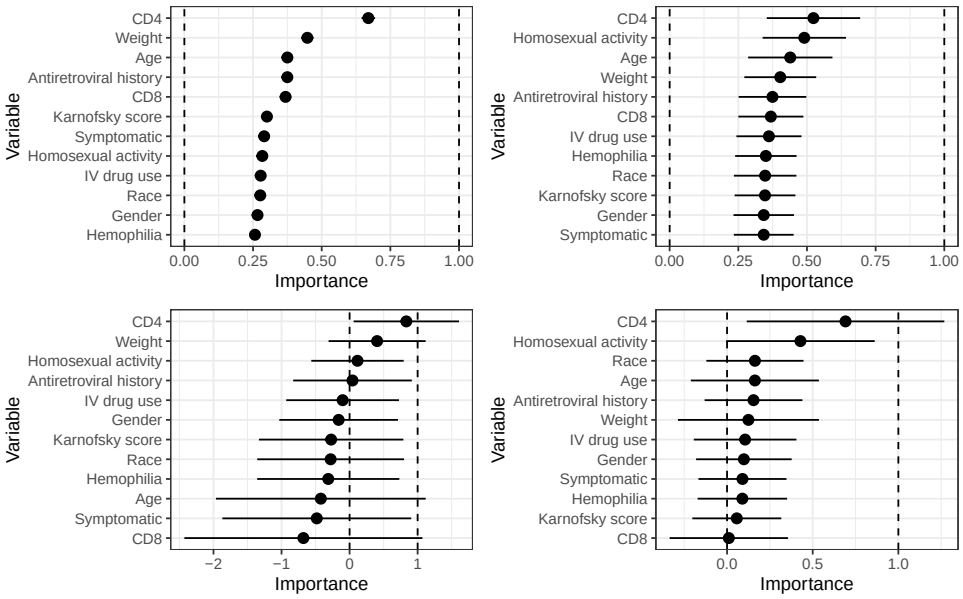

Figure 2: Results plot. Top row: no sample splitting (Algorithm 1). Bottom row: with sample splitting (Algorithm 2). Left col: T-learner (A). Right col: DR-learner (B). Black lines indicate 95% Confidence intervals.

p<0.01) where all confidence-intervals (CIs) are reported at 95% significance. CIs for the VTE were constructed using both algorithms, and we report the square root of these intervals after truncating at zero. Algorithm 2A gave a root-VTE estimate of $11.1mm^{-3}$ (CI: 0,18.3; p=0.25) and Algorithm 2B gave a root-VTE estimate of $27.5mm^{-3}$ (CI: 0, 40.2; p=0.08). These VTE estimates suggest that treatment effect homogeneity is a possible concern for our analysis. In practice, TE-VIMs are known to lie on the interval $[0, 1]$ therefore, we would recommend truncating CIs at these values. For the purposes of comparing the algorithms, however, we have chosen not do so here. The results (Figure 2) suggest CD4 T-cell count at baseline is the most important variable in determining an individuals treatment effect, with weak evidence that any other single factor is a good predictor of CATE variability. Algorithms 2A and 2B respectively estimated the TE-VIM for CD4 T-cell count to be 0.83 (CI: 0.06, 1.61; p=0.02) and 0.69 (CI: 0.12, 1.27; p=0.03).

On balance, Algorithm B provides more credible TE-VIM estimates in this applied example than Algorithm A. Both use the same pseudo-outcome estimates, however the former makes better use of the propensity score model (i.e. the fact that this is a random trial), when estimating the CATE, whereas the latter estimates the CATE based on outcome regression alone. This distinction means that we expect the CATE estimates from Algorithm B to be robust to outcome model mispecification, whereas those from Algorithm A are not. Additionally, Algorithm 2 (with sample splitting) provides more credible confidence intervals than Algorithm 1 (without sample splitting), effectively by using out-of-sample predictions to control for overfitting. As with the simulation study, we recommend Algorithm 2B for TE-VIM inference.

5 Related work and extensions

The proposed TE-VIMs complement VTEs and ATEs as an additional set of scalar summary estimands for the CATE. Fundamentally, these estimands summarise different aspects of the CATE function which are of scientific interest. Another related, and more ambitious, proposal considers the treatment effect cumulative distribution function (TE-CDF) (Levy and van der Laan, 2018), which is a curve

$$\beta(t) = Pr\{\tau(X) \leq t\}$$

parameterised by t, with $\beta(t) \in [0, 1]$ by definition. In particular, the value $\beta(0)$ is of interest since it captures the probability that an individual can be expected to have a negative treatment effect, given observed covariates X. We note that $\beta(0)$ is not the same as $Pr(Y^1 - Y^0 \leq 0)$ which suffers similar identifiability issues regarding the joint distribution of (Y^1, Y^0) as the quantity $\text{var}(Y^1 - Y^0)$ mentioned previously. Unfortunately the TE-CDF is generally not pathwise differentiable, hence Levy and van der Laan (2018) focus instead on a kernel smoothed analogue of $\beta(t)$. We remark that, provided $\tau_p > 0$, then by Chebyshev's inequality

$$\beta(0) \leq \Lambda \equiv \frac{\Theta_p}{\tau_p^2}.$$

Thus the VTE is also of scientific interest since it can be used to bound $\beta(0)$, informing investigators about the probability of negative CATEs once a positive ATE has been established. Estimation of Λ could be carried out using estimating equations estimators, as in the current work, or targeted methods (Levy et al., 2021), using the IC for Λ,

$$\frac{\{\varphi(Z) - \tau_p\}^2 - \{\varphi(Z) - \tau(X)\}^2 - \Lambda\tau_p\{2\varphi(Z) - \tau_p\}}{\tau_p^2}.$$

Such estimators are beyond the scope of the current work, though we refer the interested reader to the Appendix for a sketch of the details for the former.

The idea of treating the CATE as a statistical functional that we would like to summarise enables similar estimands to be defined in settings where one is interested in other statistical functionals. For instance, Hines

et al. (2021) propose an analogue of the CATE

$$\lambda(x) \equiv \frac{\text{cov}(A, Y|X = x)}{\text{var}(A|X = x)}$$

which is well defined even when A is a continuous exposure, and which identifies the CATE under standard causal assumptions (consistency, positivity, exchangeability) when A is binary, i.e. $\lambda(x) = \mu(1, x) - \mu(0, x)$. One might, therefore, extend the ATE, VTE, and TE-VIMs to continuous exposures by defining the estimands,

$$E\{\lambda(X)\}$$
$$\text{var}\{\lambda(X)\}$$
$$\frac{\text{var}\{\lambda(X)|X_{-s}\}}{\text{var}\{\lambda(X)\}}$$

which reduce to their CATE counterparts when A is binary. The ICs for these estimands are obtained by replacing the pseudo-outcome $\varphi(z)$ with

$$[y - \mu(x) - \lambda(x)\{a - \pi(x)\}] \frac{a - \pi(x)}{\text{var}(A|X = x)} + \lambda(x)$$

which reduces to $\varphi(z)$ when A is binary. See Appendix for details.

6 Conclusion

We propose TE-VIMs, which generalise regression-VIMs (Williamson et al., 2021a,b). These have immediate applications to the analysis of observational and clinical trial data, and provide insight into scientific questions related to treatment effect heterogeneity. Our methods complement VTE analyses, which quantify treatment effect heterogeneity (Levy et al., 2021). We derive efficient estimating equation estimators which are amenable to data-adaptive estimation of working models. These are broadly applicable, since they are not tied to a particular model or regression algorithm. We elucidate links between our estimators and regression-VIM counterparts, by interpreting our estimators in terms of pseudo-outcomes (Kennedy, 2020). We believe pseudo-outcome based approaches might generalise to other statistical functionals, where analogous pseudo-outcomes could be derived. For instance, derivative effect VIMs could be derived which reduce to TE-VIMs when exposure is binary (Hines et al., 2021), or VIMs for policy learning could be developed using double robust scores (Athey and Wager, 2021).

References

Abrevaya, J., Hsu, Y. C., and Lieli, R. P. (2015). Estimating Conditional Average Treatment Effects. *Journal of Business and Economic Statistics*, 33(4):485–505.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7353–7360.

Athey, S. and Wager, S. (2021). Policy Learning With Observational Data. *Econometrica*, 89(1):133–161.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.

Carone, M., Diaz, I., and van der Laan, M. J. (2018). Higher-Order Targeted Loss-Based Estimation. In van der Laan, M. J. and Rose, S., editors, *Targeted Learning in Data Science*, Springer Series in Statistics, chapter 26. Springer International Publishing, Cham.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. (2020). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. pages 1–27.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 78(3):655–671.

Fisher, A. and Kennedy, E. H. (2020). Visually Communicating and Teaching Intuition for Influence Functions. *The American Statistician*, pages 1–11.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *American Statistician*, 63(4):308–319.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C. (1996). Therapy in Hiv-Infected Adults With Cd4 Cell Counts. *The New England Journal of Medicine*, 335:1081–1090.

Heckman, J. J., Smith, J., and Clements, N. (1997). Making the Most out of Programme Evaluations and Social Experiments : Accounting for Heterogeneity in Programme Impacts. *Review of Economic Studies*, 64(4):487–535.

Hines, O., Diaz-Ordaz, K., and Vansteelandt, S. (2021). Parameterising the effect of a continuous exposure using average derivative effects. pages 1–25.

Hines, O., Dukes, O., Diaz-Ordaz, K., and Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 0(0):1–48.

Kallus, N., Mao, X., and Zhou, A. (2018). Interval estimation of individual-level causal effects under unobserved confounding. *arXiv*, pages 1–32.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. pages 1–35.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4156–4165.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Levy, J. and van der Laan, M. (2018). Kernel Smoothing of the Treatment Effect CDF. *arXiv*.

Levy, J., van der Laan, M., Hubbard, A., and Pirracchio, R. (2021). A fundamental measure of treatment effect heterogeneity. *Journal of Causal Inference*, 9(1):83–108.

Lu, W., Zhang, H. H., and Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22(5):493–504.

Luedtke, A. R. and van der Laan, M. J. (2016). Super-Learning of an Optimal Dynamic Treatment Rule. *International Journal of Biostatistics*, 12(1):305–332.

Newey, W. K. and Robins, J. M. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv*, pages 1–43.

Nie, X. and Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv*, (2019):1–50.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). When Some of Regression Coefficients Estimation Regressors Are Not Always Observed. *Methods*, 89(427):846–866.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186.

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

van der Laan, M. J. (2013). Targeted Learning of an Optimal Dynamic Treatment , and Statistical Inference for its Mean Outcome Targeted Learning of an Optimal Dynamic Treatment , and Statistical Inference for its Mean Outcome. *UC Berkeley Division of Biostatistics Working Paper Series*, (317):1–90.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning*, volume 27 of *Springer Series in Statistics*. Springer New York, New York, NY.

van der Laan, M. J. and Rubin, D. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1).

van der Vaart, A. W. (1998). Empirical Processes. In *Asymptotic Statistics*, pages 265–290. Cambridge University Press.

Verdinelli, I. and Wasserman, L. (2021). Decorrelated Variable Importance. pages 1–26.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021a). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2021b). A General Framework for Inference on Algorithm-Agnostic Variable Importance. *Journal of the American Statistical Association*, 0(0):1–38.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, 111(516):1548–1563.

Wright, M. N. and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1).

Zhang, L. and Janson, L. (2020). Floodgate: inference for model-free variable importance.

Zheng, W. and van der Laan, M. J. (2011). Cross-Validated Targeted Minimum-Loss-Based Estimation. In *Targeted Learning*, pages 459–474. Springer New York, New York, NY.

A Derivation of Efficient Influence Curve

To derive the ICs in (3) and (5) we adopt the formalism given in Hines et al. (2022). Specifically we let P denote the true distribution of (Y, A, X) and let $\tilde{P}$ denote a point mass at $(\tilde{y}, \tilde{a}, \tilde{x})$. We further denote the parametric submodel $P_t = t\tilde{P} + (1-t)P$ where $t \in [0,1]$ is a scalar parameter, and we let ∂_t denote an operator such that for some function of $f(t)$, $\partial_t f(t) \equiv \frac{df(t)}{dt}|_{t=0}$.

We make use of the following lemma, which we demonstrate later in the proof. Letting $g_P(X)$ denote some functional of P, then

$$\partial_t E_{P_t}\{g_{P_t}(X)|X_{-s} = x_{-s}\} = \frac{\tilde{f}(x_{-s})}{f(x_{-s})}\left[g_P(\tilde{x}) - E_P\{g_P(X)|X_{-s} = x_{-s}\}\right] + E_P\{\partial_t g_{P_t}(X)|X_{-s} = x_{-s}\} \tag{6}$$

where $\tilde{f}(.)$ and $f(.)$ denote the marginal 'densities' of X_{-s} under $\tilde{P}$ and P respectively, which are both assumed to be absolutely continuous w.r.t. to a dominating measure. In practice this expression means that for discrete X_{-s} then $f(.)$ is a probability mass function and $\tilde{f}(.)$ is an indicator function. Similarly for continuous X_{-s} then $f(.)$ is a probability density function and $\tilde{f}(.)$ is a dirac delta function. In both cases $\tilde{f}(x_{-s})$ is a probability point mass, which is zero when $\tilde{x}_{-s} \neq x_{-s}$.

It follows immediately from (6) that,

$$\partial_t E_{P_t}\{g_{P_t}(X)\} = g_P(\tilde{x}) - E_P\{g_P(X)\} + E_P\{\partial_t g_{P_t}(X)\} \tag{7}$$

By considering that

$$\text{var}_P\{g_P(X)|X_{-s} = x_{-s}\} = E_P\{g_P^2(X)|X_{-s} = x_{-s}\} - E_P\{g_P(X)|X_{-s} = x_{-s}\}^2$$

We can use (6) to show that that

$$\begin{aligned}\partial_t \text{var}_{P_t}\{g_{P_t}(X)|X_{-s} = x_{-s}\} = \frac{\tilde{f}(x_{-s})}{f(x_{-s})}\left[\{g_P(\tilde{x}) - E_P(g_P(X)|X_{-s} = x_{-s})\}^2 - \text{var}_P\{g_P(X)|X_{-s} = x_{-s}\}\right] \\ + 2\text{cov}_P\{g_P(X), \partial_t g_{P_t}(X)|X_{-s} = x_{-s}\}\end{aligned} \tag{8}$$

where $\text{cov}(A, B|C) \equiv E(\{A - E(A|C)\}B|C)$ denotes the conditional covariance. Using the results in (7) and (8), we obtain

$$\begin{aligned}\partial_t E_{P_t}\left[\text{var}_{P_t}\{g_{P_t}(X)|X_{-s}\}\right] = \{g_P(\tilde{x}) - E_P(g_P(X)|X_{-s} = \tilde{x}_{-s})\}^2 - E_P\left[\text{var}_P\{g_P(X)|X_{-s}\}\right] \\ + 2E_P\{\text{cov}_P\{g_P(X), \partial_t g_{P_t}(X)|X_{-s}\}\}\end{aligned} \tag{9}$$

Setting $g_P(X) = \tau(X)$, we use (6) and the fact that $\tau(x) = \mu(1,x) - \mu(0,x)$ to show that,

$$\partial_t g_{P_t}(x) = \frac{\tilde{f}(x)}{f(x)}\{\tilde{y} - \mu(\tilde{a}, x)\}\frac{\tilde{a} - \pi(x)}{\pi(x)\{1 - \pi(x)\}}$$

Hence, (9) implies the IC,

$$\begin{aligned}\phi_s(\tilde{z}) &= \{\tau(\tilde{x}) - \tau_s(\tilde{x})\}^2 - \Theta_s + 2\{\tau(\tilde{x}) - \tau_s(\tilde{x})\}\{\tilde{y} - \mu(\tilde{a}, \tilde{x})\}\frac{\tilde{a} - \pi(\tilde{x})}{\pi(\tilde{x})\{1 - \pi(\tilde{x})\}} \\ &= \{\tau(\tilde{x}) - \tau_s(\tilde{x})\}^2 - \Theta_s + 2\{\tau(\tilde{x}) - \tau_s(\tilde{x})\}\{\varphi(\tilde{z}) - \tau(\tilde{x})\}\end{aligned}$$

Completing the square of the expression above gives the result in (3). In replicating this proof, it is useful to note that for an arbitrary function $h(x)$

$$E_P\left\{\frac{\tilde{f}(X)}{f(X)}h(X)\right\} = h(\tilde{x})$$

Proof of Lemma in (6)

To demonstrate (6) we write the lefthand side as

$$\partial_t \int g_{P_t}(x^*) dP_{t,X_s|x_{-s}}(x_s^*) = \int g_P(x^*) \partial_t dP_{t,X_s|x_{-s}}(x_s^*) + \int \{\partial_t g_{P_t}(x^*)\} dP_{X_s|x_{-s}}(x_s^*)$$

where $dP_{t,X_s|x_{-s}}(.)$ is the conditional distribution of X_s given $X_{-s} = x_{-s}$ under the parametric submodel and $x_{-s}^* = x_{-s}$. The second integral on the righthand side recovers the final term in (6). Hence the lemma follows once we show that

$$\partial_t dP_{t,X_s|x_{-s}}(x_s^*) = \frac{\tilde{f}(x_{-s})}{f(x_{-s})} \{d\tilde{P}_{X_s}(x_s^*) - dP_{X_s|x_{-s}}(x_s^*)\}$$

To do so, let μ denote a dominating measure and write

$$\begin{aligned} dP_{t,X_s|x_{-s}}(x_s^*) &= f_{t,X_s|x_{-s}}(x_s^*) d\mu(x_s^*) \\ &= \frac{f_{t,X}(x^*)}{f_{t,X_{-s}}(x_{-s})} d\mu(x_s^*) \end{aligned}$$

where $f_{t,X}(.)$ and $f_{t,X_{-s}}(.)$ denote the marginal densities of X and X_{-s} under the parametric submodel, P_t, i.e. they are the Radon-Nikodym derivatives w.r.t. μ. Applying the quotient rule, we obtain

$$\partial_t dP_{t,X_s|x_{-s}}(x_s^*) = \frac{1}{f_{X_{-s}}(x_{-s})} \left[\partial_t f_{t,X}(x^*) - \frac{f_X(x^*)}{f_{X_{-s}}(x_{-s})} \partial_t f_{t,X_{-s}}(x_{-s}) \right] d\mu(x_s^*)$$

We now evaluate the derivative parts. Since $\partial_t P_t = \tilde{P} - P$, the marginal density derivatives will have a similar structure, as shown in the first expression below, where $f_X(.)$ and $\tilde{f}_X(.)$ denote marginal densities of X under $\tilde{P}$ and P, with likewise for X_{-s}

$$\partial_t dP_{t,X_s|x_{-s}}(x^*) = \frac{1}{f_{X_{-s}}(x_{-s})} \left[\{\tilde{f}_X(x^*) - f_X(x^*)\} - \frac{f_X(x^*)}{f_{X_{-s}}(x_{-s})} \{\tilde{f}_{X_{-s}}(x_{-s}) - f_{X_{-s}}(x_{-s})\} \right] d\mu(x_s^*)$$

Since $\tilde{P}$ is a point mass, $\tilde{f}_X(x^*) = \tilde{f}_{X_s}(x_s^*) \tilde{f}_{X_{-s}}(x_{-s}^*)$. Also $x_{-s}^* = x_{-s}$ hence,

$$\begin{aligned} \partial_t dP_{t,X_s|x_{-s}}(x^*) &= \frac{\tilde{f}_{X_{-s}}(x_{-s})}{f_{X_{-s}}(x_{-s})} \left[\tilde{f}_{X_s}(x_s^*) - \frac{f_X(x^*)}{f_{X_{-s}}(x_{-s})} \right] d\mu(x_s^*) \\ &= \frac{\tilde{f}_{X_{-s}}(x_{-s})}{f_{X_{-s}}(x_{-s})} \left[\tilde{f}_{X_s}(x_s^*) - f_{X_s|x_{-s}}(x_s^*) \right] d\mu(x_s^*) \end{aligned}$$

Thus, the result follows.

Corollary

An immediate consequence of these IC derivations is that the IC of $\text{var}\{\tau_s(X)\} = \Theta_p - \Theta_s$ is,

$$\phi_p(Z) - \phi_s(Z) = \{\varphi(Z) - \tau_p\}^2 - \{\varphi(Z) - \tau_s(X)\}^2 - \text{var}\{\tau_s(X)\}$$

This result is interesting since it holds even when Y is not independent of A given X_{-s}.

B Estimator Asymptotic Distribution

B.1 Ratio

Our goal is to consider the estimand $\Psi_s = \Theta_s/\Theta_p$ and its estimator $\hat{\Psi}_s = \hat{\Theta}_s/\hat{\Theta}_p$. Here we assume that for the numerator and denominator, we have regular asymptotically linear estimators such that,

$$\hat{\Theta}_s - \Theta_s = n^{-1}\sum_{i=1}^{n}\phi_s(z_i) + o_p(n^{-1/2})$$

Such estimators are examined in the next part of the proof. It follows by algebraic manipulations that,

$$\sqrt{n}(\hat{\Psi}_s - \Psi_s) = \frac{\Theta_p}{\hat{\Theta}_p}\left[n^{-1/2}\sum_{i=1}^{n}\Phi_s(z_i) + o_p(1)\right]$$

where $\Phi_s(z) = \{\phi_s(z) - \Psi_s\phi_p(z)\}/\Theta_p$ is the IC of Ψ_s. Next we use Slutsky's Theorem and the fact that $\hat{\Theta}_p/\Theta_p$ converges to 1 in probability, to write,

$$\lim_{n\to\infty}\sqrt{n}(\hat{\Psi}_s - \Psi_s) = \lim_{n\to\infty} n^{-1/2}\sum_{i=1}^{n}\Phi_s(z_i)$$

which gives the desired result due to the central limit theorem. We note that this set up is quite general when one considers estimands which are written as the ratio of two other estimands, such as Ψ_s in the present context.

B.2 Estimator for Θ_s

We demonstrate asymptotic regularity for the estimator $\hat{\Theta}_s$, with the result for $\hat{\Theta}_p$ following from the case $s = p$. Asymptotic regularity of Ψ_s follows using the ratio argument above.

Throughout we use superscript hat to denote functional estimators obtained from an independent sample, and we define,

$$\hat{\varphi}(z) = \{y - \hat{\mu}(a,x)\}\frac{a - \hat{\pi}(x)}{\hat{\pi}(x)\{1-\hat{\pi}(x)\}} + \hat{\mu}(1,x) - \hat{\mu}(0,x)$$
$$\hat{\phi}_s(z) = \{\hat{\varphi}(z) - \hat{\tau}_s(x)\}^2 - \{\hat{\varphi}(z) - \hat{\tau}(x)\}^2 - \hat{\Theta}_s^0$$

where $\hat{\Theta}_s^0$ is an initial plug-in estimate of Θ_s. We make the following assumptions about these functional estimators,

(A1) The differences $\tau(x) - \hat{\tau}(x)$, $\tau_s(x) - \hat{\tau}_s(x)$, and $\tau_p - \hat{\tau}_p$ are all $o_P(n^{-1/4})$ in $L_2(P)$ norm.

(A2) The propensity score and outcome estimators are 'double robust' in the sense that $\{\pi(x) - \hat{\pi}(x)\}\{\mu(a,x) - \hat{\mu}(a,x)\}$ is $o_P(n^{-1/2})$ in $L_2(P)$ norm.

(A3) The CATE difference estimates are bounded as $\{\hat{\tau}(x) - \hat{\tau}_s(x)\}^2 \leq \delta$ and $\{\hat{\tau}(x) - \hat{\tau}_p\}^2 \leq \delta$ for some $\delta < \infty$ with probability 1.

(A4) The propensity score estimates are bounded as $\epsilon \leq \hat{\pi}(x) \leq 1 - \epsilon$ for some $\epsilon > 0$ with probability 1.

(A5) There exists a P-Donsker class $\mathcal{G}_0$ such that $P(\hat{\phi}_s(.) \in \mathcal{G}_0) \to 1$ and $P(\hat{\phi}_p(.) \in \mathcal{G}_0) \to 1$.

(A6) There exists a constant $K > 0$ such that each of $\tau(x)$, $\hat{\tau}(x)$, $\hat{\tau}_s(x)$ and $\text{var}(\varphi(Z)|X = x)$ has range uniformly contained in $(-K, K)$ with probability one as $n \to \infty$. Additionally $\hat{\tau}_p \in (-K, K)$.

(A7) There exists a constant $K > 0$ such that $\text{var}(Y|X = x)$ and $\hat{\mu}(a, x)$ have range uniformly contained in $(-K, K)$ with probability one as $n \to \infty$.

Under these assumptions we show that the remainder term, R, in the expansion below is $o_P(n^{-1/2})$

$$\hat{\Theta}_s^0 - \Theta_s = -E\{\hat{\phi}_s(Z)\} + R$$

where we highlight that the expectation is conditional on the functional estimators, i.e. $\hat{\varphi}(z)$ is treated as a fixed function. We then show that

$$-E\{\hat{\phi}_s(Z)\} = n^{-1}\sum_{i=1}^{n}\phi_s(z_i) + H_n - n^{-1}\sum_{i=1}^{n}\hat{\phi}_s(z_i) \tag{10}$$

where H_n is an empirical process term, which is $o_P(n^{-1/2})$ under our assumptions. It follows therefore that for

$$\hat{\Theta}_s = \hat{\Theta}_s^0 + n^{-1}\sum_{i=1}^{n}\hat{\phi}_s(z_i)$$

$$\hat{\Theta}_s - \Theta_s = n^{-1}\sum_{i=1}^{n}\phi_s(z_i) + o_P(n^{-1/2}).$$

Formally $\hat{\Theta}_s$ is one-step plug-in bias correction estimator, but it is seen to be equivalent to the estimating equations estimator in the main text.

The remainder term

To simplify notation, in this subsection we largely omit function arguments, for example $\tau = \tau(X)$ with similar for $\hat{\tau}, \tau_s, \hat{\tau}_s, \pi, \hat{\pi}, \hat{\varphi}$. Evaluating the remainder $R \equiv E\{\hat{\phi}_s(Z) + \hat{\Theta}_s^0 - \Theta_s\}$ gives

$$R = E\left[\{\hat{\varphi} - \hat{\tau}_s\}^2 - \{\hat{\varphi} - \hat{\tau}\}^2 - \{\tau - \tau_s\}^2\right]$$

where we have used the fact that $\Theta_s = E[\{\tau - \tau_s\}^2]$. By algebraic manipulation, we write

$$R = E\left[\{\hat{\tau} - \hat{\tau}_s\}^2 - \{\tau - \tau_s\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\varphi} - \hat{\tau}\}\right]$$

We then use the identity,

$$E\left[\{\hat{\tau} - \hat{\tau}_s\}^2 - \{\tau - \tau_s\}^2\right] = E\left[\{\tau_s - \hat{\tau}_s\}^2 - \{\tau - \hat{\tau}\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\tau} - \tau\}\right]$$

to rewrite the remainder term as the sum of two error terms,

$$\begin{aligned}
R &= E\left[\{\hat{\tau} - \hat{\tau}_s\}^2 - \{\tau - \tau_s\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\varphi} - \hat{\tau}\}\right] \\
&= E\left[\{\tau_s - \hat{\tau}_s\}^2 - \{\tau - \hat{\tau}\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\tau} - \tau\} + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\varphi} - \hat{\tau}\}\right] \\
&= E\left[\{\tau_s - \hat{\tau}_s\}^2 - \{\tau - \hat{\tau}\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\varphi} - \tau\}\right] \\
&= \underbrace{E\left[\{\tau_s - \hat{\tau}_s\}^2 - \{\tau - \hat{\tau}\}^2\right]}_{\text{CATE error}} + \underbrace{2E\left[\{\hat{\tau} - \hat{\tau}_s\}r\right]}_{\text{Pseudo-outcome error}}
\end{aligned}$$

where $r = r(X)$ is defined by

$$r(x) \equiv E\left[\hat{\varphi}|X = x\right] - \tau(x)$$

This represents a pseudo-outcome error in the sense that $r(x) = E\left[\hat{\varphi} - \varphi|X = x\right]$. Splitting the remainder in to two error terms allows us to consider that the CATE error is $o_P(n^{-1/2})$ when (A1) holds. For the pseudo-outcome

error we use the Cauchy-Schwarz inequality to show that

$$E\left[\{\hat{\tau}-\hat{\tau}_s\}r\right]^2 \leq E\left[\{\hat{\tau}-\hat{\tau}_s\}^2\right] E\left[r^2\right] \leq \delta E\left[r^2\right]$$

with the second inequality following from (A3). Hence the pseudo-outcome error term is $o_P(n^{-1/2})$ if r is $o_P(n^{-1/2})$. By iterated expectation

$$r(x) = \left\{\frac{\pi(x)}{\hat{\pi}(x)} - 1\right\}\{\mu(1,x) - \hat{\mu}(1,x)\} - \left\{\frac{1-\pi(x)}{1-\hat{\pi}(x)} - 1\right\}\{\mu(0,x) - \hat{\mu}(0,x)\}$$

Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ then

$$\begin{aligned} r^2(x) &\leq 2\left\{\frac{\pi(x)}{\hat{\pi}(x)} - 1\right\}^2 \{\mu(1,x) - \hat{\mu}(1,x)\}^2 + 2\left\{\frac{1-\pi(x)}{1-\hat{\pi}(x)} - 1\right\}^2 \{\mu(0,x) - \hat{\mu}(0,x)\}^2 \\ &\leq \left(\frac{2}{\epsilon^2}\right)\{\pi(x) - \hat{\pi}(x)\}^2 \left[\{\mu(1,x) - \hat{\mu}(1,x)\}^2 + \{\mu(0,x) - \hat{\mu}(0,x)\}^2\right] \end{aligned}$$

with the second inequality following from (A4). The final expression above is $o_P(n^{-1})$ under (A2), which completes the proof that R itself is $o_P(n^{-1/2})$.

The empirical process term

In this subsection we use a common empirical processes notation, where we define linear operators P and P_n such that for some function $h(Z)$, $P\{h(Z)\} \equiv E\{h(Z)\}$ and $P_n\{h(Z)\} \equiv n^{-1}\sum_{i=1}^n h(z_i)$. Hence we write $-E\{\hat{\phi}_s(Z)\}$ as

$$(P_n - P)\{\phi_s(Z)\} + (P_n - P)\{\hat{\phi}_s(Z) - \phi_s(Z)\} - P_n\{\hat{\phi}_s(Z)\}$$

which follows from adding and subtracting $(P_n - P)\{\phi_s(Z)\}$ and $P_n\{\hat{\phi}_s(Z)\}$ to $-P\{\hat{\phi}_s(Z)\}$. This expression recovers (10) since the IC is mean zero, in the sense that $P\{\phi_s(Z)\} = 0$, and we define the empirical process term

$$H_n \equiv (P_n - P)\{\hat{\phi}_s(Z) - \phi_s(Z)\}$$

By e.g. Lemma 19.24 of van der Vaart (1998), H_n is $o_P(n^{-1/2})$ under (A5) provided that $P\left[\left\{\hat{\phi}_s(Z) - \phi_s(Z)\right\}^2\right]$ converges to zero in probability.

Start by writing,

$$\begin{aligned} \hat{\phi}_s - \phi_s =& 2(\hat{\varphi} - \varphi)(\hat{\tau} - \hat{\tau}_s) \\ &+ 2(\varphi - \tau_s)(\tau_s - \hat{\tau}_s) + (\tau_s - \hat{\tau}_s)^2 \\ &- 2(\varphi - \tau)(\tau - \hat{\tau}) - (\tau - \hat{\tau})^2 \\ &- (\hat{\Theta}_s^0 - \Theta_s) \end{aligned}$$

Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$,

$$\begin{aligned} P\left[\left\{\hat{\phi}_s(Z) - \phi_s(Z)\right\}^2\right] \leq \quad & 8P\left\{(\hat{\varphi} - \varphi)^2(\hat{\tau} - \hat{\tau}_s)^2\right] \\ &+ 2P\Big[\Big\{2(\varphi - \tau_s)(\tau_s - \hat{\tau}_s) + (\tau_s - \hat{\tau}_s)^2 \\ &\quad - 2(\varphi - \tau)(\tau - \hat{\tau}) - (\tau - \hat{\tau})^2 \\ &\quad - (\hat{\Theta}_s^0 - \Theta_s)\Big\}^2\Big] \end{aligned}$$

Letting $\hat{\Theta}_s^0 = P_n\{(\hat{\tau} - \hat{\tau}_s)^2\}$ then, in view of Theorem 1 of Williamson et al. (2021a), the second term converges to zero under (A1) and (A6). For the first of terms, we note that (A3) implies

$$P\left\{(\hat{\varphi} - \varphi)^2(\hat{\tau} - \hat{\tau}_s)^2\right\} \leq \delta P\left\{(\hat{\varphi} - \varphi)^2\right\}.$$

Similar terms to $P\left\{(\hat{\varphi} - \varphi)^2\right\}$ appear in the ATE empirical process literature. In view of Theorem 5.1 of Chernozhukov et al. (2018), this term is also converges to zero under (A2), (A4) and (A7).

Thus $H_n = o_P(n^{-1/2})$ which completes the proof.

C Additional Plots

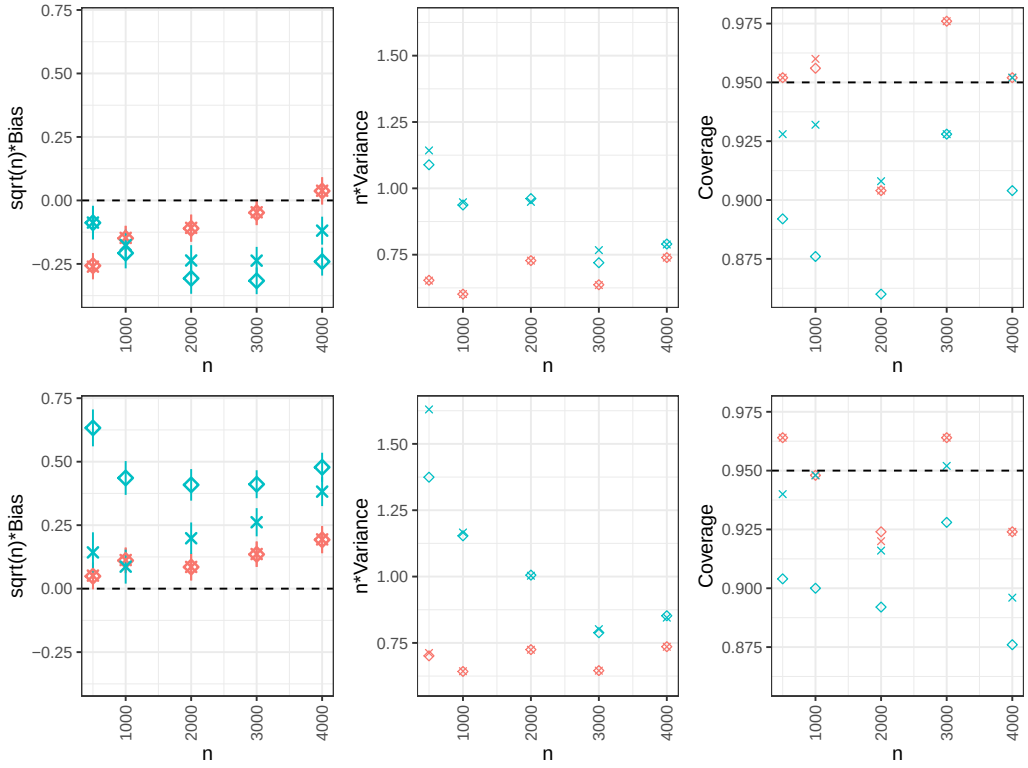

Figure 3: Bias, variance and coverage for $\hat{\Psi}_2$ using 1000 sampled datasets. Red and blue points indicate that working models are fitted using generalised additive modelling and random forests respectively. Top row of plots corresponds to Algorithm 1 (no sample splitting) and the bottom row corresponds to Algorithm 2 (sample splitting). Square and crossed points indicate that the algorithm used the T-learner and DR-learner respectively for CATE estimation.

D TE-CDF bounds

First note Chebyshev's inequality: For a variable V with mean μ and variance σ^2, for $k > 0$

$$Pr(|V - \mu| \geq k\sigma) \leq k^{-2}$$
$$Pr(V \geq \mu + k\sigma) + Pr(V \leq \mu - k\sigma) \leq k^{-2}$$

which implies the weaker inequality,

$$Pr(V \leq \mu - k\sigma) \leq k^{-2}$$

Let $\tau(X)$ be the CATE with ATE τ_p and VTE Θ_p then,

$$\beta(0) = Pr\left\{\tau(X) \leq 0\right\} = Pr\left\{\tau(X) \leq \tau_p - \left(\frac{\tau_p}{\sqrt{\Theta_p}}\right)\sqrt{\Theta_p}\right\} \leq \frac{\Theta_p}{\tau_p^2}$$

Where the inequality applies only when $\tau_p > 0$. It follows that, when the ATE is positive, the quantity on the RHS bounds $\beta(0)$ from above. The quotient rule gives that the IC (pathwise derivative) is,

$$\begin{aligned}\phi_\beta(Z) &= \frac{1}{\tau_p^2}\phi_p(Z) - 2\left(\frac{\Theta_p}{\tau_p^3}\right)\{\varphi(Z) - \tau_p\} \\ &= \frac{\{\varphi(Z) - \tau_p\}^2 - \{\varphi(Z) - \tau(X)\}^2 - \left(\frac{\Theta_p}{\tau_p^2}\right)\tau_p\{2\varphi(Z) - \tau_p\}}{\tau_p^2}\end{aligned}$$

where $\{\varphi(Z) - \tau_p\}$ is the IC of τ_p. An estimating equations estimator is that which solves

$$n^{-1}\sum_{i=1}^{n}\hat{\phi}_\beta(z_i) = 0$$

where $\hat{\phi}_\beta(z)$ is an estimate of $\phi_\beta(z)$. Therefore $\hat{\Theta}_p/\hat{\tau}_p^2$ is an estimating equations estimator where $\hat{\Theta}_p$ is the VTE estimator in the current paper and

$$\hat{\tau}_p = n^{-1}\sum_{i=1}^{n}\hat{\varphi}(z_i)$$

is the AIPW estimator of the ATE.

E IC for continuous analogue estimands

Here we use the same fomalism as in Appendix A. To derive the ICs of interest we consider the results in (7) and (9) in the setting where we set $g_P(x) = \lambda(x)$. We will show that,

$$\partial_t g_{P_t}(x) = \frac{\tilde{f}(x)}{f(x)}\{\tilde{y} - \mu(x) - \lambda(x)\{\tilde{a} - \pi(x)\}\}\frac{\tilde{a} - \pi(x)}{\text{var}(A|X = x)} \tag{11}$$

and hence, letting

$$\varphi_\lambda(z) \equiv \{y - \mu(x) - \lambda(x)\{a - \pi(x)\}\}\frac{a - \pi(x)}{\text{var}(A|X = x)} + \lambda(x)$$

then by (7) the IC of $E\{\lambda(X)\}$ is,

$$\varphi_\lambda(z) - E\{\lambda(X)\}$$

and by (9), the IC of $E\left[\text{var}\{\lambda(X)|X_{-s}\}\right]$ is,

$$\{\varphi_\lambda(z) - \lambda_s(x)\} - \{\varphi_\lambda(z) - \lambda(x)\} - E\left[\text{var}\{\lambda(X)|X_{-s}\}\right]$$

where $\lambda_s(x) = E\{\lambda(X)|X_{-s} = x_{-s}\}$. The IC for $\text{var}\{\lambda(X)\}$ follows as a special case where s includes all the observed covariates. To demonstrate (11) we first note that, by (6),

$$\begin{aligned}\partial_t \text{cov}_{P_t}(A, Y|X = x) &= \partial_t E_{P_t}\{[A - E_{P_t}(A|X)][Y - E_{P_t}(Y|X)]|X = x\} \\ &= \frac{\tilde{f}(x)}{f(x)}\left[\{\tilde{a} - \pi(x)\}\{\tilde{y} - \mu(x)\} - \text{cov}_P(A, Y|X = x)\right]\end{aligned}$$

We also obtain $\partial_t \text{var}_{P_t}(A|X = x)$ as a special case of the above expression when $Y = A$. By the quotient rule,

$$\begin{aligned}\partial_t \frac{\text{cov}_{P_t}(A, Y|X = x)}{\text{var}_{P_t}(A|X = x)} &= \frac{\partial_t \text{cov}_{P_t}(A, Y|X = x)}{\text{var}_P(A|X = x)} - \frac{\text{cov}_P(A, Y|X = x)}{\text{var}_P(A|X = x)} \frac{\partial_t \text{var}_{P_t}(A, Y|X = x)}{\text{var}_P(A|X = x)} \\ &= \frac{\tilde{f}(x)}{f(x)}\{\tilde{y} - \mu(x) - \lambda(x)\{\tilde{a} - \pi(x)\}\}\frac{\tilde{a} - \pi(x)}{\text{var}(A|X = x)}\end{aligned}$$

Thus, the desired results follow.