Appendix A
Foundations of TMLE

A.1 Asymptotic Linearity: The Functional Delta Method

Summary. An estimator of a parameter is a mapping from the data set to the parameter space. Estimators that are empirical means of a function of the unit data structure are asymptotically consistent and normally distributed due to the CLT. Such estimators are called linear in the empirical probability distribution. Most estimators are not linear, but many are approximately linear in the sense that they are linear up to a negligible (in probability) remainder term. One states that the estimator is asymptotically linear, and the relevant function of the unit data structure, centered to have mean zero, is called the influence curve of the estimator. How does one prove that an estimator is asymptotically linear? One key step is to realize that an estimator is a mapping from a possibly very large collection of empirical means of functions of the unit data structure into the parameter space. Such a collection of empirical means is called an empirical process whose behavior with respect to uniform consistency and the uniform CLT is established in empirical process theory. In this section we present succinctly that (1) a uniform central limit theorem for the vector of empirical means, combined with (2) differentiability of the estimator as a mapping from the vector of empirical means into the parameter space yields the desired asymptotic linearity. This method for establishing the asymptotic linearity and normality of the estimator is called the functional delta method (van der Vaart and Wellner 1996; Gill 1989).

Consider a sample of n i.i.d. observations $O_1, \ldots, O_n$ from a probability distribution P_0 that is known to be an element of a statistical model $\mathcal{M}$. Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be the parameter of interest, and let $\psi_0 = \Psi(P_0)$ be the true parameter value. We assume that the parameter Ψ is pathwise differentiable so that it is reasonable to assume asymptotically linear estimators of ψ_0 exist.

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_30,

Let $\mathcal{M}_{NP}$ denote a nonparametric model that includes the empirical distribution P_n of $O_1, \ldots, O_n$. Let $\hat{\Psi} : \mathcal{M}_{NP} \rightarrow \mathbb{R}^d$ be an estimator of ψ_0 that maps the empirical distribution P_n of $O_1, \ldots, O_n$ into an estimate $\hat{\Psi}(P_n)$. First, we will assume that $\hat{\Psi}(P_0) = \psi_0$ so that the estimator targets the desired target parameter ψ_0. This estimator is asymptotically linear at P_0 if $\hat{\Psi}(P_n) - \hat{\Psi}(P_0) = (P_n - P_0)IC(P_0) + o_P(1/\sqrt{n})$ for some mean zero function $IC(P_0)$ of O: i.e., $P_0 IC(P_0) = 0$. This function $IC(P_0)$ of O is called the influence curve of the estimator $\hat{\Psi}$. Notice that $(P_n - P_0)IC(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC(P_0)(O_i)$ is thus an empirical mean of mean zero i.i.d. random variables, so that the CLT immediately implies that this empirical mean is asymptotically normally distributed.

Instead of simply representing an estimator as a function of P_n, we need to be more specific by representing the estimator as a function of an empirical process $(P_n f : f \in \mathcal{F})$ for some class $\mathcal{F}$ of functions of O; that is, the estimator maps a "vector" of empirical means into the estimate. Some simple estimators, such as the sample variance, are only a function of a few empirical means, but most estimators are functions of an infinite collection of empirical means, such as whole cumulative empirical distribution functions. Similarly, $\psi_0 = \Psi(P_0 f : f \in \mathcal{F})$ is a function of the corresponding true means $P_0 = (P_0 f : f \in \mathcal{F})$. For example, in a nonparametric model $\mathcal{M}$ for a probability distribution P_0 of a multivariate Euclidean valued random variable O, ψ_0 might depend on $P_0 I(O \leq o)$ for each possible o. We will let P_0 denote the "vector" $(P_0 f : f \in \mathcal{F})$ and P_n will denote the "vector" $(P_n f : f \in \mathcal{F})$.

P_n and P_0 will be viewed as elements in a function space $\ell^\infty(\mathcal{F})$, where the latter space consists of all functions $G : \mathcal{F} \rightarrow \mathbb{R}$, endowed with the supremum norm $\|G\|_\infty = \sup_{f \in \mathcal{F}} \mid G(f) \mid$. This allows us to write $\hat{\Psi} : \ell^\infty(\mathcal{F}) \rightarrow \mathbb{R}^d$ and $\Psi : \ell^\infty(\mathcal{F}) \rightarrow \mathbb{R}^d$. With this framework in mind, we ask ourselves: Why would $\hat{\Psi}(P_n)$ be an asymptotically linear estimator of $\hat{\Psi}(P_0)$? To start with, why would $\hat{\Psi}(P_n)$ be a consistent estimator?

Formally, consistency is proven as follows. In the space $\ell^\infty(\mathcal{F})$, for a small enough class of functions $\mathcal{F}$, one has that $\|P_n - P_0\|_\infty = \sup_{f \in \mathcal{F}} \mid (P_n - P_0)f \mid$ converges to 0 in probability (as n converges to infinity). *Small enough* is measured by the entropy function of the class of functions, which is defined as the logarithm of the covering number $N(\epsilon, \mathcal{F}, \|\cdot\|)$ as a function of ϵ. The covering number $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is defined as the number of balls/spheres of size ϵ, with respect to norm $\|\cdot\|$, one needs in order to cover this set $\mathcal{F}$. Specifically, if $\sup_Q N(\epsilon \|F\|_\infty, \mathcal{F}, L_1(Q)) < \infty$ for all $\epsilon > 0$ (supremum is taken over all probability measures Q), where the function $F = \sup_{f \in \mathcal{F}} f$ is called the envelope of this class $\mathcal{F}$ of functions, then $\|P_n - P_0\|_\infty$ converges to 0 in probability. Here $L_1(Q)$ is endowed with the norm $\|f\| = \int \mid f \mid dQ$. A class $\mathcal{F}$ that satisfies this entropy condition is called a Glivenko–Cantelli class of functions. Suppose now that $\hat{\Psi} : \ell^\infty(\mathcal{F}) \rightarrow \mathbb{R}^d$ is a continuous function at P_0: i.e., if a sequence P_n converges to P_0 in $\ell^\infty(\mathcal{F})$ (i.e., with respect to the supremum norm), then $\hat{\Psi}(P_n) \rightarrow \hat{\Psi}(P_0)$. This continuity property of the mapping $\hat{\Psi}$, and the stochastic convergence $\|P_n - P_0\|_\infty \rightarrow 0$ in probability, implies now that $\hat{\Psi}(P_n)$ converges to $\hat{\Psi}(P_0)$ in probability. This is implied by the continuous mapping theorem. That is, continuity of the estimator as a mapping and $\mathcal{F}$'s being a Glivenko–Cantelli class imply the consistency of the estimator. This

typically represents the first important step proving that the estimator is also asymptotically linear, which is a stronger and much more useful statement than stating that the estimator is consistent.

For the purpose of proving asymptotic linearity, we view $G_n = \sqrt{n}(P_n - P_0)$ as a random variable in $\ell^\infty(\mathcal{F})$. In this space, for a small enough class of functions $\mathcal{F}$, one can prove that the probability distribution of G_n converges to the distribution of a Gaussian process G_0, where this Gaussian process is identified by the fact that for any finite set of functions $(f_j : j)$, $(G_0 f_j : j)$ has a multivariate normal distribution with covariances $\Sigma(f_k, f_l) = P_0(f_k - P_0 f_k)(f_l - P_0 f_l)$. Specifically, if $\int_0^\infty \sqrt{\sup_Q \log N(\epsilon, \mathcal{F}, L_2(Q))} d\epsilon < \infty$, then G_n converges in probability distribution to G_0 in $\ell^\infty(\mathcal{F})$, which is also called weak convergence and denoted by $G_n \Rightarrow_d G_0$. A class $\mathcal{F}$ for which $G_n \Rightarrow_d G_0$, such as a class satisfying this entropy condition, is called a P_0–Donsker class. Thus, whether or not $\mathcal{F}$ is a Donsker class is again determined by its entropy function. Establishing what classes of functions constitute a Donsker class is of utmost importance and is covered by empirical process theory. Beyond that the Donsker class property yields convergence of G_n to the Gaussian process G_0 in probability distribution; it also implies $\| P_n - P_0 \|_{\mathcal{F}} = O_P(1/\sqrt{n})$, and that, for any sequence $f_n \in \mathcal{F}$ so that $P_0 f_n^2 \to 0$ in probability, we have $(P_n - P_0) f_n = o_P(1/\sqrt{n})$. These types of properties are fundamental ingredients in any study of the asymptotic behavior of an estimator.

Suppose now that $\hat{\Psi} : \ell^\infty(\mathcal{F}) \to \mathbb{R}^d$ is a differentiable function at P_0. Specifically, for any sequence P_n (which can occur as a random realization of the actual empirical distribution P_n) satisfying that $\sqrt{n}(P_n - P_0) \to G_0$ in $\ell^\infty(\mathcal{F})$, we have $\sqrt{n}(\hat{\Psi}(P_n) - \hat{\Psi}(P_0)) \to \hat{\Psi}'(P_0)(G_0)$, where $\hat{\Psi}'(P_0)(G_0) = \sum_{f \in \mathcal{F}} \frac{d}{dP_0 f} \hat{\Psi}(P_0) G_0(f)$ is the directional derivative of $\hat{\Psi}$ at P_0 applied to direction G_0. Note that $\frac{d}{dP_0 f} \hat{\Psi}(P_0)$ is just the partial derivative of a function of a vector. This differentiability property of the mapping $\hat{\Psi}$ at P_0, and the stochastic convergence $G_n = \sqrt{n}(P_n - P_0) \Rightarrow_d G_0$ in distribution as random elements in $\ell^\infty(\mathcal{F})$, implies now that $\sqrt{n}(\hat{\Psi}(P_n) - \hat{\Psi}(P_0)) \Rightarrow_d \hat{\Psi}'(P_0)(G_0)$, i.e., it implies weak convergence of the standardized estimator to a d-variate normally distributed random variable $Z = \hat{\Psi}'(P_0)(G_0)$. This result is implied by the generalized continuous mapping theorem as presented in van der Vaart and Wellner (1996), applied to the functions $f_n(G_n) = \hat{\Psi}(P_0 + 1/\sqrt{n} G_n) - \hat{\Psi}(P_0)$ and its limit $f(G_0) = \hat{\Psi}'(P_0)(G_0)$. That is, an analytic differentiability property of the estimator as a mapping, and $\mathcal{F}$'s being a Donsker class imply the desired convergence in distribution of the (mean and variance-)standardized estimator to a Gaussian process. The differentiability condition and stochastic convergence of the standardized empirical process G_n, also implies

$$\sqrt{n}(\hat{\Psi}(P_n) - \hat{\Psi}(P_0)) = \hat{\Psi}'(P_0)(G_n) + o_P(1),$$

where, by linearity of the derivative $\hat{\Psi}'(P_0) : \ell^\infty(\mathcal{F}) \to \mathbb{R}^d$ and (linearity of)

$$G_n = \left(\tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n f(O_i) - P_0 f : f \in \mathcal{F} \right),$$

we have

$$\hat{\Psi}'(P_0)(G_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Psi}'(P_0)(f(O_i) - P_0 f : f).$$

Thus, $\hat{\Psi}(P_n)$ is asymptotically linear with d-dimensional influence curve

$$IC(P_0)(O) = \hat{\Psi}'(P_0)(f(O) - P_0 f : f) = \sum_{f \in \mathcal{F}} \frac{d\hat{\Psi}(P_0)}{dP_0 f}(f(O) - P_0 f).$$

Clearly, serious mathematics/functional analysis is needed to formally prove that an estimator is asymptotically linear, but it is beautiful to see that some pure analytical properties of the estimator as a mapping, and stochastic properties of the empirical process as established in empirical process theory, translate into the desired consistency and convergence in probability distribution of the standardized estimator to a normal distribution, providing a firm basis for statistical inference. In addition, the mathematics also results in the influence curve of the estimator, which has great utility in robustness analysis and estimation of the asymptotic covariance matrix of the standardized estimator $\sqrt{n}(\psi_n - \psi_0)$.

A.2 Influence Curve of an Asymptotically Linear Estimator

Summary. The functional delta method also provides us with the influence curve of the estimator. The influence curve allows robustness analysis and provides an estimator of the variance of the estimator, and thereby construction of confidence intervals and tests of null hypotheses of interest. It is a function of the unit data structure, indexed by the true probability distribution. If the true probability distribution is such that the influence curve is a nicely bounded function, then the estimator will generally behave well. This insight allows one to inspect the influence curve for necessary practical assumptions in order to have a reliable and robust estimator, without the need to formally prove mathematical theorems. One can calculate the influence curve of an estimator without formally analyzing the estimator; the influence curve is expressed in terms of the derivative of the estimator viewed as a mapping from a vector of means into the parameter space. That is, the influence curve of the estimator is expressed as a linear combination of the influence curves of the empirical means that were inputted into the estimator.

Why are we interested in calculation of the influence curve of an estimator? Given the asymptotic linearity of the estimator, the influence curve allows one to identify observations O_i that have a disproportional effect on the estimator. It teaches us under what assumptions about P_0 the influence curve $IC(P_0)$ is a nicely bounded function of O (these will be required assumptions to claim asymptotic linearity), and its covariance matrix equals the asymptotic covariance matrix of the estimator, thereby providing confidence intervals and tests of null hypotheses of interest. This will be discussed in more detail below. In addition, if one has the influence curve

IC_j of an estimator ψ_{nj} of a parameter ψ_{0j}, for each $j = 1, \ldots, d$, then one also obtains the influence curve of a function $f(\psi_{nj} : j)$ as an estimator of $f(\psi_{0j} : j)$: It is given by $\sum_j d/d\psi_{0j} f(\psi_0) IC_j$. That is, an influence curve of an estimator is a building block for calculating the influence curve of an estimator that uses this estimator as an ingredient.

Recall that P_0 denotes the "vector" $(P_0 f : f \in \mathcal{F})$ and P_n will denote the "vector" $(P_n f : f \in \mathcal{F})$. Since $\hat{\Psi}$ is a function of a vector, one needs to define directional derivatives of this function in a direction defined by a "vector" $h = (h(f) : f \in \mathcal{F})$. This directional derivative is defined as

$$\hat{\Psi}'(P_0)(h) \equiv \frac{d}{d\epsilon} \hat{\Psi}(P_0 + \epsilon h)\Big|_{\epsilon=0}.$$

Since we can think of $\hat{\Psi}$ as a function of a vector with components indexed by $f \in \mathcal{F}$, we can define a partial derivative with respect to to its fth component at $(P_0 f : f)$:

$$\frac{d\hat{\Psi}(P_0)}{dP_0 f} = \frac{d}{d\epsilon} \hat{\Psi}(P_0 + \epsilon h_f)\Big|_{\epsilon=0},$$

with $h_f(f) = 1$ and $h_f(f_1) = 0$ for $f_1 \neq f$. This partial derivative is d-dimensional, one for each component of $\hat{\Psi}$. The directional derivative in the direction of h can then be presented as a gradient (one for each of the d components of $\hat{\Psi}$) applied to vector h:

$$\begin{aligned}\hat{\Psi}'(P_0)(h) &= \frac{d}{d\epsilon} \hat{\Psi}(P_0 + \epsilon h)\Big|_{\epsilon=0} \\ &= \sum_{f \in \mathcal{F}} \frac{d\hat{\Psi}(P_0)}{dP_0 f} h(f).\end{aligned}$$

The influence curve of $\hat{\Psi}(P_n)$ under i.i.d. sampling from P_0 can be represented as the directional derivative in direction h_O defined componentwise as $h_O(f) = f(O) - P_0 f$. Note that h_O is the centered empirical process $P_{n=1} - P_0$ of one observation O, and $(P_n - P_0)(f) = 1/n \sum_{i=1}^n h_{O_i}(f)$. Thus, the influence curve can be defined as

$$\begin{aligned}IC(P_0)(O) &= \frac{d}{d\epsilon} \hat{\Psi}(P_0 + \epsilon h_O)\Big|_{\epsilon=0} \\ &= \sum_{f \in \mathcal{F}} \frac{d\hat{\Psi}(P_0)}{dP_0 f} \{f(O) - P_0 f\}.\end{aligned}$$

To summarize what we have learned, let us restate the basic delta method argument. A first-order Taylor expansion of $\hat{\Psi}$ at $P_0 = (P_0 f : f \in \mathcal{F})$ yields

$$\hat{\Psi}(P_n) - \hat{\Psi}(P_0) \approx \hat{\Psi}'(P_0)(P_n - P_0),$$

where the additional second-order term often involves differences such as $\|P_n - P_0\|^2$ or $(P_n - P_0) f_n$ with a sequence of functions f_n (depending on P_n) converging to 0 in probability. Formally, as presented previously, the functional delta method and em-

pirical process theory can be used to show that the remainder is $o_P(1/\sqrt{n})$, which is also the condition under which we can claim that the estimator $\hat{\Psi}(P_n)$ is asymptotically linear under i.i.d. sampling from P_0. Under this asymptotic linearity condition we have

$$\begin{aligned}\hat{\Psi}(P_n) - \hat{\Psi}(P_0) &= \hat{\Psi}'(P_0)(P_n - P_0) + o_P(1/\sqrt{n}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \hat{\Psi}'(P_0)(f(O_i) - P_0 f : f \in \mathcal{F}) + o_P(1/\sqrt{n}) \\ &= \frac{1}{n}\sum_{i=1}^{n} IC(P_0)(O_i) + o_P(1/\sqrt{n}).\end{aligned}$$

The influence curve of an estimator is of great importance. Firstly, it provides us with an estimator of the asymptotic covariance matrix of the estimator, which can be used for testing null hypotheses, sequential testing, sample size calculations, and the construction of confidence intervals. Specifically, $\sqrt{n}(\hat{\Psi}(P_n) - \hat{\Psi}(P_0))$ converges in distribution to $N(0, \Sigma_0)$, where $\Sigma_0 = P_0 IC(P_0) IC(P_0)^\top$. This covariance matrix can be estimated with the empirical covariance matrix of $\widehat{IC}(O_i)$, $i = 1, \ldots, n$, where $\widehat{IC}$ is an estimate of $IC(P_0)$. The asymptotically valid working model $\hat{\Psi}(P_n) \sim N(\psi_0, \hat{\Sigma})$ provides a basis for statistical inference. In addition, the influence curve values $\widehat{IC}(O_i)$, $i = 1, \ldots, n$, provide a tool for investigating the influence of one observation O_i on the estimator, and is thus also helpful for detecting outliers (robustness analysis).

Beyond its utility for assessing uncertainty and evaluating robustness, the theoretical investigation of $IC(P_0)$ allows one to determine the conditions on P_0 under which it is uniformly bounded as a function of O. These conditions will be required assumptions under which the estimator $\hat{\Psi}(P_n)$ is a reliable robust and consistent estimator of ψ_0, and these assumptions will suggest truncations or other modifications of the data so that these assumptions are met. As a consequence of this theoretical utility, a nontheoretician is able to assess quickly the situations in which the estimator is unreliable and the required conditions under which it is a trustworthy estimator. In particular, it provides insight into the amount of information, or lack of information (i.e., sparsity), in the data with respect to the target parameter ψ_0.

A.3 Computation of the Influence Curve: An Example

Summary. We present two concrete examples where we demonstrate the computation of an influence curve of an estimator. A first important step is to represent the estimator as a function of a large collection of empirical means (i.e., linear estimators), or, more generally, as a function of a collection of asymptotically linear estimators with known influence curves. Given this formulation of the estimator, for the sake of determining the influence curve of

the estimator, one should only be concerned with a first-order Taylor expansion of the estimator, viewed as a function of this collection of asymptotically linear estimators, so that one can ignore all second- and higher-order terms. In this way, the influence curve follows naturally as a linear combination of the influence curves of the inputted estimators. That is, the influence curve equals the (functional) derivative of the estimator applied to the vector of influence curves of the inputted estimators.

We demonstrate the computation of an influence curve of an estimator. Let $O = (W, A, Y) \sim P_0$, and let $\Psi(P) = E_P[E_P(Y \mid W, A = 1) - E_P(Y \mid W, A = 0)]$ be the parameter of interest, $O_1, \ldots, O_n$ be n i.i.d. observations of O, and P_n be the empirical probability distribution.

Influence curve of MLE based on parametric model. Suppose that we use a parametric model $\{\bar{Q}_\beta : \beta\}$ for $\bar{Q}_0$, where $\bar{Q}_0(A, W) = E_0(Y \mid W, A)$. For example, $\bar{Q}_\beta(A, W) = \beta^\top(A, W)$ is a main term linear regression model. Let β_n be an estimator of β according to this parametric model. Such an estimator β_n solves an estimating equation such as $P_n D_{\beta_n} = 0$. For example, if $L(\bar{Q})$ is a loss function for $\bar{Q}_0$, such as $L(\bar{Q})(O) = (Y - \bar{Q}(W, A))^2$, and $\beta_n = \arg\min_\beta P_n L(\bar{Q}_\beta)$, then $D(\beta) = \frac{d}{d\beta} L(\bar{Q}_\beta)$. Let β_0 be the limit of β_n satisfying $P_0 D_{\beta_0} = 0$. Then, under regularity conditions, it follows that $\beta_n - \beta_0 = (P_n - P_0) IC_{\beta_0} + o_P(1/\sqrt{n})$, where $IC_{\beta_0} = c_0^{-1} D_{\beta_0}$, with $c_0 = -\frac{d}{d\beta_0} P_0 D_{\beta_0}$.

We will also use the notation $\tilde{Q}_\beta(W) \equiv \bar{Q}_\beta(W, 1) - \bar{Q}_\beta(W, 0)$. Let $\psi_0 = P_{W,0} \tilde{Q}_{\beta_0}$ be the target parameter of interest. Thus, if the parametric model is correctly specified, we have $\psi_0 = E_0[E_0(Y \mid W, A = 1) - E_0(Y \mid W, A = 0)]$. Let $\psi_n = P_{W,n} \tilde{Q}_{\beta_n}$ be the estimator of ψ_0. We wish to determine the influence curve of ψ_n as an estimator of ψ_0. We have

$$\begin{aligned}
\psi_n - \psi_0 &= P_{W,n} \tilde{Q}_{\beta_n} - P_{W,0} \tilde{Q}_{\beta_0} \\
&= (P_{W,n} - P_{W,0}) \tilde{Q}_{\beta_0} + P_{W,n} \{\tilde{Q}_{\beta_n} - \tilde{Q}_{\beta_0}\} \\
&= (P_{W,n} - P_{W,0}) \tilde{Q}_{\beta_0} + P_{W,0} \{\tilde{Q}_{\beta_n} - \tilde{Q}_{\beta_0}\} + (P_{W,n} - P_{W,0}) \{\tilde{Q}_{\beta_n} - \tilde{Q}_{\beta_0}\}.
\end{aligned}$$

The last term is a second-order term and can therefore be ignored for the purpose of calculating of an influence curve. The second term can be approximated by applying first-order Taylor expansions in β and the asymptotic linearity of β_n:

$$\begin{aligned}
P_{W,0} \{\tilde{Q}_{\beta_n} - \tilde{Q}_{\beta_0}\} &\approx \left\{ P_{W,0} \frac{d}{d\beta_0} \tilde{Q}_{\beta_0} \right\}^\top (\beta_n - \beta_0) \\
&\approx (P_n - P_0) \left\{ P_{W,0} \frac{d}{d\beta_0} \tilde{Q}_{\beta_0} \right\}^\top IC_{\beta_0}.
\end{aligned}$$

We can conclude that $\psi_n - \psi_0 \approx (P_n - P_0) IC$ with influence curve

$$IC = \tilde{Q}_{\beta_0} - \psi_0 + P_{W,0} \left\{ \tfrac{d}{d\beta_0} \tilde{Q}_{\beta_0} \right\}^\top IC_{\beta_0}.$$

Influence curve of nonparametric MLE. Let us now consider a nonparametric estimator $\psi_n = P_{W,n}\tilde{Q}_n$ of $\psi_0 = P_{W,0}\tilde{Q}_0$, where $\tilde{Q}_n(W) = \bar{Q}_n(W,1) - \bar{Q}_n(W,0)$ and $\tilde{Q}_0(W) = E_0(Y \mid W, A=1) - E_0(Y \mid W, A=0)$. It is assumed that W is discrete. Note that

$$\bar{Q}_n(w,a) = \textstyle\sum_y y\,(P_n f_{y,w,a}/P_n f_{w,a}),$$

with $f_{w,a}(O) = I(W=w, A=a)$ and $f_{y,w,a}(O) = I(W=w, A=a, Y=Y)$. We will focus on deriving the influence curve IC_1 of $\psi_n(1) = P_{W,n}\bar{Q}_{1,n}$ as an estimator of $\psi_0(1) = P_W\bar{Q}_{1,0}$, where $\bar{Q}_{1,0}(W) = E_0(Y \mid W, A=1)$. Since $\psi_n = \psi_n(1) - \psi_n(0)$, this will yield the influence curve $IC_1 - IC_0$ of ψ_n. Note that $\psi_n(1)$ can be represented as a function Φ of $(P_n f : f \in \mathcal{F})$ with $\mathcal{F} = \{f_{w,a}, f_{y,w,a} : w, a, y\}$. Thus $\psi_n(1) = \Phi(P_n) = \Phi(P_n f : f)$. The functional delta method teaches us that we can use the first order linear approximation $\Phi(P_n) - \Phi(P_0) = \Phi'(P_0)(P_n - P_0)$, where $\Phi'(P_0) = (\frac{d}{dP_0 f}\Phi(P_0 f : f) : f)$ and $(P_n - P_0) = ((P_n - P_0)f : f)$. In particular, it follows that the influence curve of $\psi_n(1) = \Phi(P_n)$ as an estimator of $\psi_0(1)$ is given by

$$IC_1 = \left(\tfrac{d\Phi(P_0)}{dP_0 f} : f\right)(P_{n=1} - P_0),$$

where $P_{n=1} = (f(O) - P_0 f : f \in \mathcal{F})$ is the empirical process based on a single observation $O = (W, A, Y)$. One can also carry out this process of determining the linear approximation in a stepwise fashion. Firstly, we linearize $\Phi(P_n)$ in terms of $P_{W,n} - P_{W,0}$ and $\bar{Q}_{1n} - \bar{Q}_{1,0}$:

$$P_{W,n}\bar{Q}_{1,n} - P_{W,0}\bar{Q}_{1,0} \approx (P_{W,n} - P_{W,0})\bar{Q}_{1,0} + P_{W,0}(\bar{Q}_{1,n} - \bar{Q}_{1,0}).$$

Note that the first term equals the empirical mean of $\bar{Q}_{1,0}(W) - \psi_0(1)$. Secondly, we linearize the latter term:

$$\bar{Q}_{1,n}(w) - \bar{Q}_{1,0}(w) \approx \textstyle\sum_y y \frac{1}{P_0 f_{w,1}}(P_n - P_0)f_{y,w,1} - \frac{P_0 f_{y,w,1}}{P_0^2 f_{w,1}}(P_n - P_0)f_{w,1}.$$

Thus,

$$\begin{aligned}\sum_w P_{W,0}(w)(\bar{Q}_{1,n} - \bar{Q}_{1,0})(w) &\approx (P_n - P_0)\left\{\sum_w P_{W,0}(w)\sum_y y\left\{\frac{1}{P_0 f_{w,1}}f_{y,w,1} - \frac{P_0 f_{y,w,1}}{P_0^2 f_{w,1}}f_{w,1}\right\}\right\} \\ &\equiv (P_n - P_0)IC_1'.\end{aligned}$$

Since $f_{y,w,1} = I(Y=y, W=w, A=a)$ and $f_{w,1} = I(W=w, A=1)$, the integrals/sums over w, y simplify:

$$\begin{aligned}IC_1' &= \frac{P_0(W)}{P_0(W, A=1)}YI(A=1) - P_0(W)\sum_y y\frac{P_0(y,W,1)}{P_0^2(W,1)}I(A=1) \\ &= \frac{I(A=1)}{g_0(1 \mid W)}Y - \frac{I(A=1)}{g_0(1 \mid W)}E_0(Y \mid W, A=1) \\ &= \frac{I(A=1)}{g_0(1 \mid W)}(Y - E_0(Y \mid W, A=1)),\end{aligned}$$

where we used the notation $g_0(1 \mid W) = P_0(A = 1 \mid W)$. Thus, the influence curve IC_1 of $\psi_n(1)$ is given by $\bar{Q}_{1,0}(W) - \psi_0(1) + IC_1'(O)$. The above proof also yields the analog influence curve IC_0 of $\psi_n(0)$. As a consequence, we have shown that the influence curve IC of $\psi_n = \psi_n(1) - \psi_n(0)$ is given by $IC_1 - IC_0$, which can be represented as

$$IC(P_0)(O) = \left\{ \frac{I(A=1)}{g_0(1|W)} - \frac{I(A=0)}{g_0(0|W)} \right\} (Y - \bar{Q}_0(W, A)) + \tilde{Q}_0(W) - \Psi(Q_0).$$

We note that this influence curve equals the efficient influence curve of ψ_0 in the nonparametric model for P_0. Because the nonparametric MLE is asymptotically linear with influence curve equal to the efficient influence curve, by definition, it is an efficient estimator. Computing the influence curve of the (nonparametric) MLE of a target parameter in a nonparametric or semiparametric model, ignoring second-order terms and assuming the data structure is discrete, is a general tool for deriving the efficient influence curve of the target parameter. The resulting expression will have a natural analog for the general (e.g., continuously valued) data structure since each continuous data structure can be approximated by discrete data structures, and this generalized expression will then be the efficient influence curve.

Estimation of the influence curve of the nonparametric MLE, or any other efficient estimator, requires an estimator of the treatment mechanism. If one used a data-adaptive machine learning algorithm to estimate $E_0(Y \mid A, W)$ instead of a nonparametric MLE, and claimed that the resulting MLE-based estimator of the target parameter are still unbiased *enough*, then one would claim that it was asymptotically linear with influence curve equal to the efficient influence curve. Thus, estimation of the influence curve requires estimation of the treatment mechanism, again. Overall, one can conclude that statistical inference based on an MLE in a nonparametric model still requires implicit or explicit estimation of the treatment mechanism. From this point of view, the TMLE is not asking for more than what a nontargeted MLE already requires; the TMLE just utilizes the estimator of the treatment mechanism to target the estimator so that the desired asymptotic linearity is a more reasonable assumption.

A.4 Cramer–Rao Lower Bound

Summary. We prove that the influence curve of an asymptotically linear estimator of a statistical parameter that also satisfies a regularity property has a variance that is larger than the variance of the canonical gradient of the pathwise derivative of the statistical parameter. As a consequence, we can state that an estimator is optimal/efficient among all such asymptotically linear estimators if and only if its influence curve equals the canonical gradient. This explains why the latter is also-called the efficient influence curve. This result is implied by the more general convolution theorem for regular esti-

mators (Bickel et al. 1997), but it provides a self-contained understanding of efficiency theory for asymptotically linear estimators. Given the efficiency theory, one should always be highly motivated to determine the efficient influence curve of the target parameter. Indeed, it provides the ingredient for the construction of an efficient substitution estimator, such as the TMLE.

We provide a basic understanding of the result that an estimator is efficient among regular estimators if and only if it is asymptotically linear with influence curve equal to the efficient influence curve. We will prove a result stating that an asymptotically linear estimator at P_0 that maintains low negligible bias in local neighborhoods of P_0 has an influence curve that equals a gradient of the pathwise derivative. As a consequence, the best estimator among such asymptotically linear estimators is the one with an influence curve equal to the canonical gradient of the pathwise derivative.

Let $O \sim P_0 \in \mathcal{M}$, and let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be the target parameter. For each $P \in \mathcal{M}$, consider a class of parametric models $\{P_h(\epsilon) : \epsilon\} \subset \mathcal{M}$ through a $P \in \mathcal{M}$ at $\epsilon = 0$, indexed by an h in an index set $\mathcal{H}$, and with score $S(h) = \frac{d}{d\epsilon} \log P_h(\epsilon)\Big|_{\epsilon=0}$ at $\epsilon = 0$. It is assumed that these scores $\{S_h : h \in \mathcal{H}\}$ are an element of the Hilbert space $L_0^2(P)$ of mean 0 functions of O, endowed with the inner product $\langle f, g\rangle = Pfg$, the covariance operator. The set of scores generated by this class of parametric models spans a linear subspace of $L_0^2(P)$, and by taking the closure of this linear subspace we obtain the tangent space $T(P) \subset L_0^2(P)$ at P, which is itself a Hilbert space. We state that $\Psi : \mathcal{M} \to \mathbb{R}^d$ is pathwise differentiable at P if there exists a D^* in the tangent space $T(P) \subset L_0^2(P)$ at P so that for each of these submodels through P we have

$$\frac{d}{d\epsilon} \Psi(P(\epsilon))\Big|_{\epsilon=0} = PD^*(P)S.$$

This inner product representation of the derivative can be expected since 1) the left-hand side is linear in $\frac{d}{d\epsilon} P(\epsilon)\Big|_{\epsilon=0}$, so that at this fixed P, it should also be linear in the score $S = \frac{d}{d\epsilon} P(\epsilon)\Big|_{\epsilon=0} / P$, and 2) by the Riesz representation theorem, a bounded linear operator on a Hilbert space (i.e., $T(P)$) can be represented as an inner product as above. One refers to $D^*(P)$ as the canonical gradient of the pathwise derivative of $\Psi : \mathcal{M} \to \mathbb{R}^d$ at P. If the inner-product representation applies for a $D(P) \in L_0^2(P)$, then such a $D(P)$ is called a gradient. We note that for any $D^\perp$ in the orthogonal complement of the tangent space, $T(P)$, $D^*(P) + D^\perp$ is a gradient: $P(D^*(P) + D^\perp)S = PD^*(P)S$ for all $S \in T(P)$. Since a gradient has to yield the same pathwise derivative on the tangent space $T(P)$ as the canonical gradient, it follows that any gradient can be represented as $D^*(P) + D^\perp$. This shows that the set of all gradients is any function in $L_0^2(P)$ whose projection on $T(P)$ equals the canonical gradient.

Thus the canonical gradient is the unique gradient of the pathwise derivative that is an element of the tangent space, and it is also the gradient that has the smallest variance among all gradients:

$$P\{D^*(P)(O) + D^\perp\}^2 = P\{D^*(P)\}^2 + P\{D^\perp\}^2 \geq P\{D^*(P)\}^2.$$

Since an asymptotically linear estimator is efficient if and only if its influence curve equals the canonical gradient, the canonical gradient is also called the efficient influence curve.

Let us formalize the latter statement. Consider an estimator $\hat{\Psi} : \mathcal{M}_{NP} \to \mathbb{R}^d$ that maps the empirical distribution of the data set $O_1, \ldots, O_n$ into the parameter space of Ψ. Suppose the estimator is asymptotically linear under i.i.d. sampling from P with influence curve $IC(P)$:

$$\hat{\Psi}(P_n) - \Psi(P) = (P_n - P)IC(P) + R(P_n, P),$$

where $R(P_n, P) = o_P(1/\sqrt{n})$. We now argue that, for any parametric model $\{P(\epsilon) : \epsilon\}$ (one of the models used in the pathwise derivative) this linear approximation in the data P_n should still be valid under i.i.d. sampling from $P(\epsilon_n)$ with $\epsilon_n = 1/\sqrt{n}$. That is, if $P(\epsilon_n)_n$ is the empirical distribution of n i.i.d. observations from $P(\epsilon_n)$, then

$$\hat{\Psi}(P(\epsilon_n)_n) - \Psi(P) = (P(\epsilon_n)_n - P)IC(P) + R(P(\epsilon_n)_n, P),$$

where $R(P(\epsilon_n)_n, P) = o_P(1/\sqrt{n})$. This is indeed expected since $P(\epsilon_n)_n - P = (P(\epsilon_n)_n - P(\epsilon_n)) + (P(\epsilon_n) - P)$, and, if $\epsilon_n = 1/\sqrt{n}$, then $\sup_{f \in \mathcal{F}} \mid (P(\epsilon_n)_n - P(\epsilon_n))f \mid = O_P(1/\sqrt{n})$ for a Donsker class $\mathcal{F}$, while also $\| P(\epsilon_n) - P \|_{\mathcal{F}} = O(1/\sqrt{n})$. For example, to use simplistic notation, if $R(P_n, P) = \|P_n - P\|^2$, then $R(P(\epsilon_n)_n, P) = \|P(\epsilon_n)_n - P\|^2$ will converge to zero in probability at same rate as $R(P_n, P)$. As a consequence, it is reasonable to state that the linear approximation of $\hat{\Psi}(P_n)$ in the data P_n at P also holds up under sampling from $P(\epsilon_n)$, that is, under i.i.d. sampling from $P(\epsilon_n)$, we have

$$\hat{\Psi}(P(\epsilon_n)_n) - \Psi(P) = \frac{1}{n} \sum_{i=1}^n IC(P)(O_i) + R_n,$$

where $R_n = o_P(1/\sqrt{n})$. Suppose we now also require that the estimator $\hat{\Psi}(P(\epsilon_n)_n)$ has a bias that is $o(1/\sqrt{n})$ under i.i.d. sampling from $P(\epsilon_n)$ for $\epsilon_n = 1/\sqrt{n}$ in the sense that

$$\frac{E_{P(\epsilon_n)} \hat{\Psi}(P(\epsilon_n)_n) - \Psi(P(\epsilon_n))}{\epsilon_n} \to 0 \text{ as } n \to \infty, \tag{A.1}$$

$$\frac{E_{P(\epsilon_n)} R_n}{\epsilon_n} \to 0, \tag{A.2}$$

and these two statements need to hold for each of the parametric submodels. We will now show that this requirement implies that $IC(P) = D^*(P) + D^\perp$ for some $D^\perp \perp T(P)$, i.e., it implies that $IC(P)$ is a gradient of the pathwise derivative at P. This then also proves that an estimator, among the class of asymptotically linear estimators that are locally uniformly unbiased in the above sense, is efficient if and only if $IC(P) = D^*(P)$. Note that if (A.1) does not hold for a particular parametric submodel, then for this submodel $\sqrt{n}(\hat{\Psi}(P(\epsilon_n)_n) - \Psi(P(\epsilon_n)))$ will converge to a normal distribution with a bias term, possibly even a bias term of infinite magnitude, so that statistical inference based on the CLT under sampling from $P(\epsilon_n)$ will not be valid.

The fact that this "negligible bias" requirement under $P(\epsilon_n)$ implies that $IC(P)$ is a gradient is shown as follows. Firstly, substitute $\hat{\Psi}(P(\epsilon_n)_n) = \Psi(P) + P(\epsilon_n)_n IC(P) + R_n$ in (A.1) to obtain

$$\begin{aligned}\frac{E_{P(\epsilon_n)}\hat{\Psi}(P(\epsilon_n)_n) - \Psi(P(\epsilon_n))}{\epsilon_n} &= \frac{P(\epsilon_n)IC(P)}{\epsilon_n} - \frac{\Psi(P(\epsilon_n)) - \Psi(P)}{\epsilon_n} + E_{P(\epsilon_n)}\frac{R_n}{\epsilon_n} \\ &= P\frac{P(\epsilon_n) - P}{\epsilon_n P}IC(P) - \{PD^*(P)S + o(1)\} + o(1),\end{aligned}$$

where we used (A.2) and the definition of the pathwise differentiability of Ψ. Note that

$$P\frac{P(\epsilon_n)-P}{\epsilon_n P}IC(P) = P\, IC(P)S + o(1).$$

Thus, we obtain that the limit for $\epsilon_n \to 0$ (i.e., A.1) equals:

$$P\{IC(P) - D^*(P)\}S \text{ for each } S \in T(P).$$

By assumption (A.1), this limit must equal zero for all S. This proves the statement that the projection of the influence curve onto the tangent space at P is unique, and equals the canonical gradient: $\Pi(IC(P) \mid T(P)) = D^*(P)$. We will state what we just proved as a theorem.

Theorem A.1. *Let $O \sim P \in \mathcal{M}$, and $\Psi : \mathcal{M} \to \mathbb{R}^d$. Consider a class of parametric submodels $\{P_h(\epsilon) : \epsilon\} \subset \mathcal{M}$ through $P \in \mathcal{M}$ at $\epsilon = 0$, with score $S(h) = \frac{d}{d\epsilon}\log P_h(\epsilon)\big|_{\epsilon=0}$ at $\epsilon = 0$, indexed by h in an index set $\mathcal{H}$. Let $T(P) \subset L_0^2(P)$ be the tangent space at P of this class of parametric submodels. Assume that Ψ is pathwise differentiable at P with respect to this class of parametric submodels; we have that there exists a D^* in the tangent space $T(P) \subset L_0^2(P)$ at P so that for each of these submodels through P, we have*

$$\frac{d}{d\epsilon}\Psi(P(\epsilon))\Big|_{\epsilon=0} = PD^*(P)S.$$

Here we use the notation $Pf = \int f(o)dP(o)$.

Consider an estimator $\hat{\Psi} : \mathcal{M}_{NP} \to \mathbb{R}^d$ that maps an empirical distribution P_n of $O_1, \ldots, O_n \sim P$ (for any P) into $\mathbb{R}^d$. Assume that for each of the above submodels $\{P_h(\epsilon) : \epsilon\}$, $h \in \mathcal{H}$, under i.i.d. sampling from $P(\epsilon_n)$ (suppressing h) with $\epsilon_n = 1/\sqrt{n}$, we have

$$\hat{\Psi}(P(\epsilon_n)_n) - \Psi(P) = P(\epsilon_n)_n IC(P) + R_n,$$

with

$$\frac{E_{P(\epsilon_n)}\hat{\Psi}(P(\epsilon)_n) - \Psi(P(\epsilon_n))}{\epsilon_n} \to 0 \text{ as } n \to \infty, \tag{A.3}$$

$$E_{P(\epsilon_n)}R_n = o(\epsilon_n). \tag{A.4}$$

Here $P(\epsilon)_n$ is the empirical distribution of an i.i.d. sample of size n from $P(\epsilon)$. Then the projection of $IC(P)$ onto $T(P)$ in the Hilbert space $L_0^2(P)$ equals the canonical gradient $D^*(P)$:

$$\Pi(IC(P) \mid T(P)) = D^*(P).$$

In particular,

$$VAR_P\{IC(P)(O)\} \geq VAR_P\{D^*(P)(O)\}.$$

Why the variance of the efficient influence curve is a generalized Cramer–Rao lower bound: An informal explanation. If one assumes a parametric model $\{P(\epsilon) : \epsilon\}$ so that $O \sim P(\epsilon_0)$, and one wishes to estimate $f(\epsilon_0) = \Psi(P(\epsilon_0))$, where we let $\epsilon_0 = 0$ (unknown to the user), then the Cramer–Rao lower bound for the variance of an unbiased estimator of $f(\epsilon_0)$ is given by

$$\frac{\left(\frac{d}{d\epsilon_0} f(\epsilon_0)\right)^2}{PS^2} = \frac{P^2 D^*(P) S}{PS^2}.$$

Here S is the score of $P(\epsilon)$ at $\epsilon = 0$. Since each such parametric models is a submodel, it makes sense to define as the Cramer–Rao bound for the actual model $\mathcal{M}$, the worst-case bound obtained by selecting the hardest among a class of parametric submodels. For most models $\mathcal{M}$, if one has univariate submodels with scores S_1 and S_2, then one can also construct a submodel with two parameters whose score is any linear combination of S_1 and S_2. In that case, the worst-case bound is given by

$$\sup_{S \in T(P)} \frac{P^2 D^*(P) S}{PS^2}.$$

By the Cauchy–Schwarz inequality, it follows that this supremum is attained at $S = D^*(P)$ and it equals $PD^*(P)^2$, the variance of the efficient influence curve.

A.5 Invariance of Statistical Properties

Summary. Given the statistical model and target parameter, we emphasize that statistical properties, such as the pathwise derivative of the target parameter, the gradient and canonical gradient/efficient influence curve, and robustness of this efficient influence curve, are invariant to additional nontestable assumptions that do not change the statistical model. This insight is useful, since it allows one to borrow from previously obtained statistical results.

Consider a random variable (U, X), a model $\mathcal{M}^F$ for its distribution $P_{U,X}$, and a mapping from $P_{U,X} \in \mathcal{M}^F$ into a probability distribution $P_O(P_{U,X})$ for an observed data structure O (e.g., $O = \Phi(U, X)$ for some mapping Φ). We use the notation (U, X) for the underlying random variable because we have used this notation to describe the random variables modeled by an SCM. However, (U, X) can denote any underlying random variable; for example, U might be the censoring variable, X a full data structure, and $O = \Phi(U, X)$ the observed data structure. Let

$\mathcal{M} = \{P_O(P_{U,X}) : P_{U,X} \in \mathcal{M}^F\}$ be the corresponding observed data statistical model. Consider a random variable (U^*, X^*), a model $\mathcal{M}^{F*}$ for its distribution P_{U^*,X^*}, and a mapping from $P_{U^*,X^*} \in \mathcal{M}^{F*}$ into a probability distribution $P_O(P_{U^*,X^*})$ for the observed data structure O. Let $\mathcal{M}^* = \{P_O(P_{U^*,X^*}) : P_{U^*,X^*} \in \mathcal{M}^{F*}\}$ be the corresponding observed data statistical model. If we write $\mathcal{M}$ we also refer to the actual model assumptions coded by the parameterization $P_O(P_{U,X})$ and underlying model $\mathcal{M}^F$, and similarly for $\mathcal{M}^*$. Although models $\mathcal{M}$ and $\mathcal{M}^*$ can be very different in their underlying assumptions, we assume that their statistical models are identical: $\mathcal{M}^* = \mathcal{M}$.

In general, if one proves statistical properties that concern the probability distribution of the observed data structure O under one of these models, it will also apply to the other model. This might sound too trivial to even mention, but it is a useful fact. In order to be concrete, let us consider a number of scenarios in which we apply this invariance principle.

Suppose that one has proven the following double robustness results for an observed data estimating function $(Q, g) \to D(Q, g)$, with respect to nuisance parameters $Q(P)$ and $g(P)$, based on model $\mathcal{M}$ [i.e., the proof might have used the parameterization $P_O(P_{U,X})$ and assumptions $P_{U,X} \in \mathcal{M}^F$]. For each $P \in \mathcal{M}$ and $Q \in \mathcal{Q}$, there is a set $\mathcal{G}(Q, P)$ of distributions such that $E_P D(Q, g)(O) = 0$ for $g \in \mathcal{G}(Q, P)$. Then, we also have for each $P \in \mathcal{M}^*$ and $Q \in \mathcal{Q}$, $E_P D(Q, g)(O) = 0$ for $g \in \mathcal{G}(Q, P)$.

For example, consider the CAR missing-data model for $O = (W, A, Y = Y_A) \sim P_0$ with $X = (Y_0, Y_1, W)$ nonparametrically modeled, and the treatment assignment mechanism $g_0(a \mid X) = P_0(A = a \mid X) = P_0(A = a \mid W)$ only assumed to satisfy CAR. Let $D^*(P_0) = D^*(Q_0, g_0)$ be the efficient influence curve of the target parameter $\Psi(P_0) = E_0[E_0(Y \mid A = 1, W) - E(Y \mid A = 0, W)]$, with Q_0 representing $E_0(Y \mid A, W)$ and the marginal distribution of W, so that both Q_0, g_0 represent parameters of P_0. Suppose that one has proven a desired double robustness result in this CAR missing-data model such as $P_0 D^*(Q, g) = 0$ if either $Q = Q_0$ or $g = g_0$ (van der Laan and Robins 2003). Then this same double robustness result applies to other causal underlying models that result in the same nonparametric observed data model, including the pure statistical model for $O = (W, A, Y) \sim P_0$ that makes no assumptions at all.

One possible example is that model $\mathcal{M}$ makes stronger assumptions than $\mathcal{M}^*$ (e.g., $\mathcal{M}^F \subset \mathcal{M}^{F*}$ or $\mathcal{M}^*$ is a pure statistical model); apparently, proving a result under stronger assumptions implies the result under weaker assumptions if these stronger assumptions do not shrink the statistical model. For example, if one proves statistical properties under the causal model for a longitudinal data structure $(\bar{A}, \bar{L}_A)$ that assumes a strong SRA, $A(t) \perp L_{\bar{A}(t-1),\underline{a}(t)} \mid \bar{A}(t-1), \bar{L}(t)$, then these same statistical properties apply under the causal model that assumes the weaker randomization with respect to the Y outcome, where the Y outcome is included in the L process. We know the latter randomization assumption is sufficient for identification of marginal distributions of Y_a (but not for L_a), but it might prevent us from doing calculations and engaging in reasoning that feels natural to us.

Similarly, suppose that one has proven that, given a loss function $L(Q)$, for each $P \in \mathcal{M}$ and $Q(P) = \arg\min_{Q \in \mathcal{Q}} PL(Q)$, we have that $\Psi(Q(P))$ is a desired

number $\Phi(P)$, and this proof used model $\mathcal{M}$. Then this also implies that for each $P \in \mathcal{M}^*$, and $Q(P) = \arg\min_{Q \in \mathcal{Q}} PL(Q)$, we have that $\Psi(Q(P)) = \Phi(P)$. For example, consider the CAR missing-data model for $O = (W, A, Y = Y_A)$ again. Consider the loss function $L(Q) = -\log Q(Y \mid A)g_0(A)/g_0(A \mid X)$. Using the CAR censored-data model, it follows that this is a valid IPTW loss function for $Q_0(y \mid a) = P_0(Y(a) = y)$, i.e., for the marginal distribution of the counterfactual Y_a for each a: $Q_0 = \arg\min_Q E_0 L(Q)$. As a consequence, we can identify the additive causal effect as follows: $E_0\{Y_1 - Y_0\} = \sum_y y\{Q_0(y \mid a = 1) - Q_0(y \mid a = 0)\}$. The right-hand side defines now a mapping Φ from P_0 into a desired value $\Phi(P_0) = \Psi(P_0) = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$. Now, suppose we are not willing to make any causal assumptions. Then we still have $\Phi(P_0) = \Psi(P_0)$, allowing one to construct estimators based on the mapping Φ, even though Q_0 no longer represents a counterfactual distribution.

Another application of this invariance principle is the following. Suppose we have shown that $D(P)$ is a gradient of a parameter $\Psi : \mathcal{M} \to \mathbb{R}$ at $P \in \mathcal{M}$. That is, for each one-dimensional submodel $\{P(\epsilon) : \epsilon\}$ through P at $\epsilon = 0$, $\frac{d}{d\epsilon}\Psi(P(\epsilon))\big|_{\epsilon=0} = E_P D(P)S$, where S is the score of $P(\epsilon)$ at $\epsilon = 0$. Suppose this proof (seemingly) relied on the model $\mathcal{M}$, including its nontestable assumptions. Then, $D(P)$ is also a gradient of the parameter $\Psi : \mathcal{M}^* \to \mathbb{R}$. Thus, the strategy for deriving the set of all gradients or the canonical gradient might be to assume various nontestable assumptions, such as representing the observed data structure as a CAR censored data structure. If one determines a gradient or canonical gradient under these assumptions, then one has also determined a gradient or canonical gradient under another model that maps into the same statistical model for the distribution of O.

This strategy allows us to utilize results from the literature. In particular, van der Laan and Robins (2003) present a theory for determining the class of gradients of a target parameter for a statistical model for the observed data structure under the assumption that the observed data structure is a function of a full-data random variable and censoring variable and that the conditional distribution of the censoring variable, given the full-data random variable, satisfies coarsening at random. In particular, it shows how to map the gradients of an identifiable target parameter in the full-data model into the gradients for the corresponding target parameter in the observed data model. Because of the invariance principle, this provides us with the set of gradients and the canonical gradient of the target parameter of the observed data distribution for the statistical model for O, regardless of the nontestable underlying assumptions one makes. This insight allows us to borrow from the rich literature on CAR censored-data models. For example, the censored-data literature provides us with IPCW estimating functions that were shown to be gradients of the pathwise derivative in the observed data model in which the censoring mechanism is assumed known. By projecting a gradient on the tangent space of the model one obtains the canonical gradient, so that these IPTW estimating functions can be used to derive the canonical gradient. If one does not assume that the observed data are represented as a censored-data structure, it does not make sense to use such naming or talk about full data estimating functions, but nevertheless, we can still temporarily

move ourselves into this world to make progress by building on theory developed in that world, and use IPCW estimating functions (that were developed and presented in this CAR censored data world) as functions of the observed data that are gradients of the statistical target parameter.

In the CAR censored-data model for $O = (W, A, Y = Y_A)$, with full data structure $X = (W, Y_0, Y_1)$ nonparametrically modelled, and g_0 known, $\{I(A = 1) - I(A = 0)\}/g_0(A \mid X)Y - \psi_0$ represents an IPTW estimating function for ψ_0 (which is verified by showing that its conditional expectation, given X, yields a gradient in the full data model), so that we can also use this gradient in the pure statistical model, but one now represents g_0 as the conditional distribution of A, given W.

Suppose that one assumes a time-ordering $A \rightarrow W \rightarrow Y$, where A might be gender, W intermediate variables, and Y a final outcome of interest. One assumes an SCM according to this time ordering, and one assumes that (A, W) is randomized. Suppose that one is interested in a direct effect of gender on salary, controlled by the intermediate variables W, of the type $E_0 \sum_w \{Y(1, w) - Y(0, w)\} P_0(w \mid A = 0)$. This parameter can be identified from the observed data as

$$E_0 \sum_w \{Y(1, w) - Y(0, w)\} P_0(w \mid A = 0) = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W) \mid A = 0].$$

Suppose one has derived the canonical gradient of this target parameter of the observed data distribution, as presented in van der Laan (2010c).

Consider now a new causal model in which one wishes to estimate the effect of treatment among the nontreated. That is, one assumes the time ordering $A \rightarrow W \rightarrow Y$, where W are baseline covariates, A is a binary treatment, and Y is a final outcome of interest. One assumes an SCM according to this ordering, and one assumes that A is randomized, conditional on W. Suppose that one is interested in the causal effect of treatment among the nontreated $E_0(Y_1 - Y_0 \mid A = 0)$. This parameter is identified from the observed data $O = (W, A, Y)$ as

$$E_0(Y_1 - Y_0 \mid A = 0) = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W) \mid A = 0].$$

Even though the two causal models and the causal quantities are very different, the observed data structures are identical, the statistical models are identical, and the statistical target parameters are identical. As a consequence, we now also have the canonical gradient for the latter causal model and target parameter.

Finally, let us consider another kind of application of this invariance of statistical properties under varying nontestable assumptions. Suppose that, given a parameter $\Psi^{F*} : \mathcal{M}^{F*} \rightarrow \mathbb{R}$, one has shown that $\Psi^{F*}(P_{U^*,X^*}) = \Psi^*(P_O(P_{U^*,X^*}))$ for each $P_{U^*,X^*} \in \mathcal{M}^{F*}$ for some mapping $\Psi^* : \mathcal{M} \rightarrow \mathbb{R}$. In other words, one has established the identifiability of the causal parameter Ψ^{F*} in model $\mathcal{M}^*$ through the statistical parameter Ψ^*. Suppose that one has proposed a new mapping $\hat{\Psi} : \mathcal{M} \rightarrow \mathbb{R}$ and one is able to show: for each $P \in \mathcal{M}$, $\hat{\Psi}(P) = \Psi^*(P)$. Thus the two parameters $\hat{\Psi}$ and Ψ^* defined in the two models $\mathcal{M}$ and $\mathcal{M}^*$ are identical as statistical parameters. Then, $\Psi^{F*}(P_{U^*,X^*}) = \hat{\Psi}(P_O(P_{U^*,X^*}))$ for each $P_{U^*,X^*} \in \mathcal{M}^{F*}$: i.e., one also has the

identifiability of the causal parameter Ψ^{F*} in model $\mathcal{M}^*$ through the statistical parameter $\hat{\Psi}$. This can be used to establish that a particular estimator is indeed valid for estimating a desired causal effect; one shows that it statistically agrees with the mapping Ψ^* that came out of an original identifiability result.

A.6 Targeted Minimum-Loss-Based Estimation

Summary. We present a natural generalization of targeted maximum likelihood estimation, demonstrating that TMLE requires specifying an appropriate loss function and fluctuation working model so that the derivative at zero fluctuation of the loss yields the desired estimating function, such as the efficient influence curve.

Let O be the observed data structure, and let P_0 be its probability distribution. In addition, let $\mathcal{M}$ be the statistical model for P_0, and let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be a pathwise differentiable d-dimensional parameter. One observes n i.i.d. copies $O_1, \ldots, O_n$ of O and one wishes to construct an estimator of $\Psi(P_0)$. Suppose that $Q_0 = Q(P_0)$ represents a parameter $Q : \mathcal{M} \to \mathcal{Q}$ so that for some Ψ^1 we have $\Psi(P) = \Psi^1(Q(P))$ for all $P \in \mathcal{M}$. Let $\mathcal{Q} = \{Q(P) : P \in \mathcal{M}\}$ be the parameter space for Q. For notational convenience, we will use notation $\Psi(P)$ and $\Psi(Q)$ interchangeably. We wish to construct a substitution estimator $\Psi(Q_n^*)$ of ψ_0 obtained by substitution of an estimator Q_n^* of Q_0 into the parameter mapping Ψ. Let $L(Q)$ be a loss function for Q_0 so that $Q_0 = \arg\min_{Q \in \mathcal{Q}} P_0 L(Q)$. We will allow this loss function to be indexed by a nuisance parameter: $L(Q) = L_{\eta_0}(Q)$ for some unknown nuisance parameter $\eta_0 = \Gamma(P_0)$. Given an estimator of η_0, one can use loss-based (e.g., super) learning to construct an estimator Q_n^0 of Q_0 (e.g., van der Laan and Dudoit 2003).

We are not satisfied with a good estimator of Q_0. Instead, we wish to construct an updated estimator Q_n^* so that Q_n^* and η_n solve a particular estimating equation $P_n D(Q_n^*, \eta_n) = 0$ for a user-supplied target-parameter-specific estimating function $D(Q, \eta)$. The choice of this estimating function D is tailored so that solving this equation implies good properties for the substitution estimator $\Psi(Q_n^*)$ of ψ_0. For example, $D(Q_0, \eta_0)$ might be the canonical gradient (i.e., efficient influence curve) of the pathwise derivative of Ψ at P_0, and solving the efficient influence curve estimating equation is known to imply that $\Psi(Q_n^*)$ is asymptotically linear with influence curve equal to the efficient influence curve under appropriate conditions.

For any possible (Q, η), let $\{Q_\eta(\epsilon) : \epsilon\} \subset \mathcal{Q}$ be a submodel with a finite-dimensional parameter ϵ that contains Q at $\epsilon = 0$, typically indexed by η, that satisfies the following local condition at $\epsilon = 0$:

$$\frac{d}{d\epsilon} L_\eta(Q_\eta(\epsilon))\Big|_{\epsilon=0} = D(Q, \eta).$$

The targeted minimum loss based estimator (also TMLE) is now defined by the following iterative algorithm. Start with initial estimator Q_n^0, and for $k = 1, \ldots,$

define $Q_n^k = Q_{n,\eta_n}^{k-1}(\epsilon_n^k)$, where $\epsilon_n^k = \arg\min_\epsilon P_n L_{\eta_n}(Q_{n,\eta_n}^{k-1}(\epsilon))$, and stop at step k when $\epsilon_n^k \approx 0$. If $\epsilon_n^k = 0$ and it is a local minima at an interior point, then it follows that the final update $Q_n^* = Q_n^k$ solves $0 = P_n D(Q_n^*, \eta_n)$. The substitution estimator $\Psi(Q_n^*)$ is the targeted minimum-loss-based estimator of ψ_0.

Suppose $\frac{d}{d\epsilon_j} L_\eta(Q_\eta(\epsilon))\Big|_{\epsilon=0} = D_j(Q,\eta)$, while $D(Q,\eta) = \sum_j D_j(Q,\eta)$. One can also select an ordering for $(\epsilon_1, \ldots, \epsilon_J)$ (e.g., starting at ϵ_J and going backwards) and, according to this ordering, iteratively carry out the update step $Q_n^k = Q_{n,\eta_n}^{k-1}(\epsilon_n^k)$, but where ϵ_n^k is now obtained by minimizing $P_n L_{\eta_n}(Q_{n,\eta_n}^{k-1}(\epsilon))$ only over the next ϵ-component according to the ordering of the ϵ-components, using the previous value ϵ_n^{k-1} for all other components. The next ϵ-component of the last ϵ-component in this ordering is defined as the first ϵ-component in the ordering, so that one keeps circling through all ϵ-components. At convergence, we have that ϵ_n solves $P_n D_j(Q_n^*, \eta_n) = 0$ for all j, and thus $P_n D(Q_n^*, \eta_n) = 0$ as well.

The asymptotic linearity of $\Psi(Q_n^*)$ can now be based on the fact that Q_n^* solves this estimating equation, and on statistical properties of (Q_n^*, g_n) as an estimator of Q_0, g_0 (see the asymptotic linearity theorem in Appendix A.1). By selecting a loss function for Q_0 (e.g., log-likelihood loss function), and a fluctuation working model so that the linear span of the derivative of $L(Q(\epsilon))$ at $\epsilon = 0$ includes the components of the efficient influence curve of Ψ at P, one obtains a TMLE that is asymptotically efficient under appropriate conditions.

A.7 Efficient Influence Curve for Longitudinal Data Structures

Summary. We demonstrate how one calculates the efficient influence curve of a target parameter of interest for the longitudinal data structures covered in this book. The canonical gradient is a projection of an initial gradient onto the tangent space generated by scores of parametric submodels through the data-generating distribution, where this projection is carried out in the Hilbert space of mean zero functions of the unit data structure O, endowed with an inner product equal to the covariance operator. We show that a factorization of the likelihood of the unit data structure yields an orthogonal decomposition of this tangent space, and thereby of this projection as a sum of projections on orthogonal subtangent spaces. We show that these projections on the subtangent spaces can be represented in terms of conditional expectations.

Consider a set of variables $O = (O(j) : j)$ and corresponding parent variables $(Pa(O(j)) : j)$ and suppose that the probability distribution of O is given by

$$P_0(O) = \prod_j P_0(O(j) \mid Pa(O(j))).$$

For example, O could be represented by an ordered sequence of variables, and the jth variable $O(j)$ in the sequence has corresponding parents $Pa(O(j)) = \bar{O}(j-1) =$

$O(1), \ldots, O(j-1)$. Typically, $O(0)$ represents the baseline covariates. Consider the statistical model $\mathcal{M}$ implied by all such possible probability distributions, without putting any constraints on $P_0(O(j) \mid Pa(O(j)))$. In the special case where $Pa(O(j)) = \bar{O}(j-1)$, this statistical model is completely nonparametric. We use the short-hand notation $P_{O(j)}$ for the conditional distribution of $O(j)$, given $Pa(O(j))$, under P.

Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be the d-dimensional statistical parameter of interest, so that $\Psi(P_0)$ is the target parameter value we wish to learn from the data consisting of n i.i.d. draws $O_1, \ldots, O_n$ of the random variable O. We wish to determine the efficient influence curve/canonical gradient $D^*(P)$ of the pathwise derivative of Ψ at P. The canonical gradient $D^*(P)$ equals the projection $\Pi(D(P) \mid T(P))$ of an initial gradient $D(P)$ of the pathwise derivative of Ψ at P onto the tangent space $T(P)$.

Consider a rich class of submodels $P(\epsilon)$ that only vary $P_{O(j)|Pa(O(j))}$, and denote the tangent space generated by this class of submodels by $T_{O(j)}(P)$. We can do this for each of the factors indexed by $j = 1, \ldots, J$. The resulting union of parametric submodels generates a tangent space $T(P)$ at P, given by the sum space $T(P) = \sum_j T_{O(j)}(P)$. One can also observe that by adding parametric submodels that simultaneously fluctuate multiple factors in the factorization of P one generates scores that are sums of scores that are thus still contained in this sum space $\sum_j T_{O(j)}(P)$. These subtangent spaces $T_{O(j)}(P)$ are pairwise orthogonal due to the factorization of the probability density in terms of these conditional distributions $P_{O(j)}$. This shows that $T(P) = \sum_j T_{O(j)}(P)$ is an orthogonal decomposition in subspaces. The tangent space of $P_{O(j)}$ can be generated by the following parametric submodels through $P_{O(j)}$: $P_{O(j),\epsilon} = (1 + \epsilon S(O(j) \mid Pa(O(j)))P_{O(j)}$, where S is any function of $(O(j), Pa(O(j)))$ with conditional mean zero, given $Pa(O(j))$. The scores S of these parametric fluctuations at $\epsilon = 0$ generate the tangent space $T_{O(j)}(P)$, so that we have

$$T_{O(j)}(P) = \{S(O(j), Pa(O(j)) : E_P(S \mid Pa(O(j)) = 0\} \subset L_0^2(P).$$

The projection of a function D onto $T_{O(j)}(P)$ is obtained by first projecting onto all functions of $(O(j), Pa(O(j)))$, which is given by $E(D \mid O(j), Pa(O(j)))$, and subsequently projecting this projection onto all functions which also have conditional mean zero, given $Pa(O(j))$, resulting in

$$\Pi(D(P) \mid T_{O(j)}(P)) = E(D(P) \mid O(j), Pa(O(j))) - E(D(P) \mid Pa(O(j))).$$

The reader can directly verify the validity of this formula by proving that it is an element of the space $T_{O(j)}(P)$, and that $D(P)$ minus this projection is orthogonal to $T_{O(j)}(P)$. We conclude that, if $D(P)$ is a gradient of the pathwise derivative of Ψ at P, then the canonical gradient $D^*(P)$ can be determined as

$$D^*(P) = \sum_j \{E(D(P) \mid O(j), Pa(O(j))) - E(D(P) \mid Pa(O(j)))\}.$$

The efficient influence curves presented in this book have all been determined with this recipe.

One might now wonder, how do I obtain this initial gradient? One approach is to come up with an ad hoc estimator of ψ_0 that is regular and asymptotically linear, and derive its influence curve. As shown in Theorem A.1, this influence curve is now a gradient and can thus be used as initial gradient $D(P)$. We can make this more specific by adding some additional structure by defining $\mathcal{I}$ as the index set that identifies the intervention nodes $(O(j) : j \in \mathcal{I})$, and by assuming that $\Psi(P_0)$ is only a function of $P_{O(j),0}$ for $j \in \mathcal{I}^c$. Let $Q_0 = \prod_{j\in\mathcal{I}^c} P_{O(j),0}$, and let $g_0 = \prod_{j\in\mathcal{I}} P_{O(j),0}$, and note that $P_0 = Q_0 g_0$. We will use the notation $\Psi(Q_0)$ for the target parameter to stress that it only depends on P_0 through Q_0. For example, if our target parameter represents a parameter of the g-computation formula for the counterfactual distribution of O under interventions on the intervention nodes, then, indeed, $\Psi(P_0)$ is only a function of these conditional distributions $P_{O(j),0}$ with $j \in \mathcal{I}^c$.

In this case, we can use the following trick to determine a gradient. We consider the submodel $\mathcal{M}(g_0)$ of $\mathcal{M}$, which assumes that Q_0 is unspecified as in the actual model $\mathcal{M}$ but that g_0 is known. Due to the factorization of $P_0 = Q_0 g_0$ and that Ψ is only a function of P_0 through Q_0, it follows that the canonical gradient of $\Psi : \mathcal{M} \to \mathbb{R}^d$ is identical to the canonical gradient of $\Psi : \mathcal{M}(g_0) \to \mathbb{R}^d$. As a consequence, we can now act as if $\mathcal{M}(g_0)$ is the statistical model and determine the canonical gradient in this smaller model, which will then also equal the desired canonical gradient for the actual model $\mathcal{M}$. Note that the tangent space, say, $T_{g_0}(P)$ of $\mathcal{M}(g_0)$, is now only generated by fluctuations of the conditional distributions of $O(j)$, $j \in \mathcal{I}^c$ since g_0 is known. Thus the tangent space $T_{g_0}(P)$ at $P \in \mathcal{M}(g_0)$ for this model $\mathcal{M}(g_0)$ is given by $T_{g_0}(P) = \sum_{j\in\mathcal{I}^c} T_{O(j)}(P)$. In this model, with g_0 known, it is often easy to determine an ad hoc regular asymptotically linear estimator of ψ_0 that utilizes the known g_0, so that its influence curve gives us the desired initial gradient $D(P)$, whose projection onto the tangent space $T_{g_0}(P)$ of model $\mathcal{M}(g_0)$ yields the canonical gradient. Specifically, relying on the invariance of statistical properties under varying nontestable assumptions and the theory presented in van der Laan and Robins (2003), one can construct the inverse probability of censoring (g_0)-weighted estimators of the type

$$\psi_{IPCW,n} = \frac{1}{n}\sum_{i=1}^n \frac{h(O_i)}{g_0(O_i)},$$

where h is chosen such that $E_0 h(O)/g_0(O) = \psi_0$. This estimator has influence curve $D(P) = h/g_0 - \psi_0$. Or, more generally, one might define $\psi_{IPCW,n}$ as a solution of an estimating equation $P_n h(\psi)/g_0 = 0$, so that the influence curve of $\psi_{IPCW,n}$ is given by

$$D(P_0) = -\left[\frac{d}{d\psi_0} P_0 h(\psi_0)/g_0\right]^{-1} \frac{h(\psi_0)}{g_0}.$$

One can also apply the inverse weighting to the different components of a sum representation $h = \sum_j h_j$, so that $\psi_{IPCW,n}$ is defined as the solution of

$$0 = \frac{1}{n}\sum_{i=1}^n \sum_j \frac{h_j(\bar{O}_i(j),\psi)}{\prod_{l\in\mathcal{I},l<j} g_{O(l),0}}(\bar{O}_i(l)),$$

and its influence curve is the standardized version of $\sum_j h_j(\psi_0)/\prod_{l\in\mathcal{I};l<j} g_{O(l),0}$.

Thus, for example, if we take the initial gradient of the form $D(P) = h/g_0 - \psi_0$, then we obtain the following explicit representation of the canonical gradient:

$$D^*(P) = E\left(\frac{h}{g_0} \mid O(0)\right) - \psi + \sum_{j \in \mathcal{I}^c} E\left(\frac{h}{g_0} \mid O(j), Pa(O(j))\right) - E\left(\frac{h}{g_0} \mid Pa(O(j))\right).$$

Let us assume that $Pa(O(j))$ includes all the intervention nodes $(O(l) : l \in \mathcal{I}(j))$, where, for notational convenience, we defined $\mathcal{I}(j) = \{l \in \mathcal{I} : l < j\}$. In this case, these conditional expectations can always be factorized as $1/\prod_{l \in \mathcal{I}(j)} g_{O(l),0}$, times a conditional expectation that is determined by the Q_0-factor only. For example, if $Pa(O(j)) = \bar{O}(j-1)$, then

$$E_0(h/g_0 \mid O(j), Pa(O(j))) = \frac{1}{\prod_{l \in \mathcal{I}(j)} g_{O(l),0}} E_{Q_0}\left(\sum_{o(l): l \in \mathcal{I}(j)^c} h \mid O(j), Pa(O(j))\right),$$

where the sum sums up over all possible realizations of the intervention nodes $(O(l) : l \in \mathcal{I}(j)^c)$ after j. We used notation $\mathcal{I}(j)^c = \{l \in \mathcal{I} : l \notin \mathcal{I}(j)\}$ for the intervention nodes with index larger than j. The conditional expectation corresponds with taking the expectation of h with respect to to the counterfactual distribution of $(O(l) : l > j)$ under which the intervention nodes $l \in \mathcal{I}(j)^c$ are set to a fixed value $o(l)$, summed up over all possible realizations of the intervention nodes. Thus,

$$E_{Q_0}\left(\sum_{o(l): l \in \mathcal{I}(j)^c} h \mid O(j), Pa(O(j))\right) = \sum_{o(l): l > j} h(o) \prod_{l > j} Q_{0,O(l)}(\bar{o}(l)) \prod_{l \in \mathcal{I}(j)^c} g^*_{O(l)}(\bar{o}(l)),$$

where $g^*_{O(l)}(\bar{o}(l)) = 1$. The same applies to the conditional expectation of h/g_0, given $Pa(L(j))$. Thus, this yields the following explicit representation of the canonical gradient:

$$\begin{aligned} D^*(P_0)(O) = {} & E_{Q_0}\left(\sum_{o(l): l \in \mathcal{I}} h \mid O(0)\right) - \Psi(Q_0) \\ & + \sum_{j \in \mathcal{I}^c} \frac{1}{\prod_{l \in \mathcal{I}(j)} g_{O(l),0}} \left\{ E_{Q_0}\left(\sum_{o(l): l \in \mathcal{I}(j)^c} h \mid O(j), Pa(O(j)\right) \right. \\ & \left. - E_{Q_0}\left(\sum_{o(l): l \in \mathcal{I}(j)^c} h \mid Pa(O(j)\right)\right\}. \end{aligned} \tag{A.5}$$

Representation of the efficient influence curve based on factorization of the likelihood in terms of binary conditional distributions. One can always represent a longitudinal data structure in terms of an ordered sequence of binary random variables. Let $L(k)$ be a particular variable measured as part of the longitudinal data structure. One can decompose $L(k)$ in terms of binaries $(L(k,l) : l = 1, \ldots, l_k)$. For example, for a univariate continuous covariate $L(k)$, we could partition its range, and set $L(k,l) = I(L(k) = (a_l, a_{l+1}])$ for the lth interval $(a_l, a_{l+1}]$ in the partitioning of its range. If $L(k)$ consists of several univariate covariates, then one first orders these covariates, discretizes each of them, and creates corresponding indicator variables. It follows that, given a certain ordering for all the binary variables coding $L(k)$, one can parameterize the conditional distribution of $L(k)$, given $Pa(L(k))$, as

$$P_{L(k)}(L(k)) = \prod_l P_{L(k,l)}(L(k,l) \mid Pa(L(k,l))),$$

where $Pa(L(k,l)) = (Pa(L(k)), L(k,1), \ldots, L(k,l-1))$.

Therefore, we will assume that this kind of preprocessing of the data has been carried out, so that $O(l)$ for $l \geq 1$ are all binary random variables, while the baseline covariates $O(0)$ can be a vector of continuous and discrete covariates. Since $O(l)$ is a binary variable, we can represent any function S of $O(l)$ and $(O(l))$ with conditional mean zero, given $Pa(O(l))$, as

$$S = \{S(1 \mid Pa(O(l)) - S(0 \mid Pa(O(l))\}\{O(l) - P_{O(l)}(1 \mid Pa(O(l)))\}.$$

The projection of an initial gradient D (or any other function) onto $T_{O(l)}(P)$ is given by

$$D^*_{O(l)}(P) = H^*_{O(l}(Pa(O(l)))\{O(l) - P_{O(l)}(1 \mid Pa(O(l))\},$$

where the term $H^*_{O(l)}$ in front of the residual of $O(l)$ plays a crucial role as the clever covariate in the TMLE algorithm, and is given by

$$H^*_{O(l}(Pa(O(l))) \equiv E(D \mid O(l) = 1, Pa(O(l))) - E(D \mid O(l) = 0, Pa(O(l))).$$

The efficient influence curve can thus be represented as

$$D^*(P) = \sum_{k \in I^c} D^*_{O(k)}(P) = \sum_{k \in I^c} H^*_{O(k)}(O(k) - P_{O(k)}(1)),$$

where we used short-hand notation. Above we showed that

$$H^*_{O(k)}(Q,g) = H^*_{O(k),g} H^*_{O(k),Q}$$

factorizes in a g-factor $H^*_{O(k),g}$ and a Q-factor $H^*_{O(k)}(Q)$ that equals a conditional expectation with respect to $\prod_{l>k} Q_{O(l)}$. It is of interest to note that $H^*_{O(k)}(Q)$ only depends on Q through the "future" factors $Q_{O(l)}$, $l > k$; This monotonicity property of H^* allows a particular convenient closed-form implementation of the TMLE, presented in detail in next section (van der Laan 2010a,b).

A.8 Factorization in Terms of Binary Conditional Distributions

Summary. The efficient influence curve allows one to construct an efficient estimator of the target parameter. The TMLE uses the efficient influence curve to define a parametric submodel through an initial estimator of the data-generating distribution, whose parametric maximum likelihood estimator defines the targeted update of the initial estimator, and the iterative application of this updating step defines the TMLE. We present such a TMLE for general longitudinal data structures, based on a factorization of the observed data density in terms of conditional distributions of binary random variables. The targeted updates can be computed based on standard logistic regression software.

The TMLE is defined by a choice of loss function $L(Q)$ and a submodel $\{Q(\epsilon) : \epsilon\}$ through Q at $\epsilon = 0$, where we will require that $\langle D^*(Q,g)\rangle \subset \langle < \frac{d}{d\epsilon} L(Q(\epsilon))\big|_{\epsilon=0} \rangle >$, where $D^*(Q,g)$ is the efficient influence curve at $P = Qg$, and, for a function $f = (f_1, \ldots, f_K)$, $\langle f \rangle$ denotes the linear span of the components of the function f in $L_0^2(P)$. We consider two such choices and thereby two types of TMLE that will be asymptotically equivalent, since both will solve the efficient influence curve estimating equation.

TMLE I. Let

$$P(\epsilon) = \prod_{j \in \mathcal{I}^c} Q(\epsilon)_{O(j)} \prod_{j \in \mathcal{I}} g_{O(j)},$$

where for $j \geq 1$

$$\text{logit}Q(\epsilon)_{O(j)}(1) = \text{logit}Q_{O(j)}(1) + \epsilon_j H^*_{O(j)}(Q,g)$$

is a logistic regression model using the logit of $P_{O(j)}(1 \mid Pa(O(j)))$ as offset, and $H^*_{O(j)} = H^*_{O(j)}(Q,g)$ as d-dimensional covariate of the same dimension as the target parameter ψ_0. Here ϵ_j is a subvector of ϵ. In addition, the fluctuation model $Q(\epsilon)_{O(0)}$ for the distribution of the baseline covariates $O(0)$ is chosen to have a score of $D^*_{O(0)}(Q)$ with respect to ϵ_0. Let $L(Q) = -\log Q$ be the log-likelihood loss function. Indeed, the score $\frac{d}{d\epsilon} \log P(\epsilon)$ of $P(\epsilon)$ at $\epsilon = 0$ equals or spans (if ϵ is multivariate) $D^*(P)$, and the score of $Q_{O(j)}(\epsilon_j)$ at $\epsilon_j = 0$ equals $D^*_{O(j)}(P)$. Thus, $\langle D^*(Q,g)\rangle \subset \langle < \frac{d}{d\epsilon} L(Q(\epsilon))\big|_{\epsilon=0} \rangle$.

Note that the maximum likelihood estimator of ϵ_j for a given initial $P = Qg$ can be determined with univariate logistic regression software regressing the binary $O(j)$ on the clever covariate $H^*_{O(j)}(Q,g)$, using the initial as offset. If one uses a common ϵ, i.e., $\epsilon_j = \epsilon$ for $j > 0$, then one can fit this single ϵ by regressing the binary outcome $O(j)$ on the clever covariate $H^*_{O(j)}(Q,g)$ based on a *pooled* data set, so that all j-specific logistic regressions with common parameter ϵ are fit in one run.

Consider an initial estimator $P_n^0 = Q_n^0 g_n^0$ of P_0, where $Q^0_{O(0),n}$ is the empirical distribution of the baseline covariates $O_i(0)$, $i = 1, \ldots, n$. We will use a separate ϵ_0 for the fluctutation of $Q^0_{O(0),n}$, and it will always equal 0, so that this empirical distribution will not be updated. Given the loss function $L(P) = -\log Q(P)$, we determine

$$\epsilon_n^1 = \arg\min_\epsilon P_n L(P_n^0(\epsilon)).$$

This results in the first-step TMLE $P_n^1 = P_n^0(\epsilon_n^1)$. This updating process $P_n^k = P_n^{k-1}(\epsilon_n^k)$, $k = 1, \ldots, K$, is iterated to convergence defined by $\epsilon_n^k \approx 0$. The final update is the TMLE of P_0 and is denoted by $P_n^* = Q_n^* g_n^0$. We note that g_n^0 is not updated in this process due to the fluctuation working model only allowing fluctuations of Q_n^0. The TMLE of ψ_0 is now defined as the substitution estimator $\Psi(P_n^*) = \Psi(Q_n^*)$.

One may use a separate ϵ_j for each factor $Q^0_{O(j),n}$, $j \geq 1$. These maximum likelihood estimators of ϵ_j can again be determined with logistic linear regression software, as remarked above. Importantly, one may determine the maximum likelihood estimators of these fluctuation parameters recursively, starting with the last factor and working backward to the first factor, always using the most recent update of

the estimator of Q_0. In principle, one would start over at the last factor after having finished the update of the first factor and iterate this updating process until convergence. However, it follows that, with this recursive algorithm, the TMLE requires only one update per factor, and thereby converges in one step (representing one round from the last factor to the first factor, always using the most recent update for Q_0) and exists in analytic form. This algorithm is introduced and presented in detail in van der Laan (2010a). The convergence in one round is due to the above-mentioned monotonicity property of $H^*_{O(k)}(Q)$ with respect to its dependence on the Q-factors $Q_{O(k)}$ (van der Laan 2010a,b).

TMLE II. Let

$$P(\epsilon) = \prod_{j \in \mathcal{I}^c} Q(\epsilon)_{O(j)} \prod_{j \in \mathcal{I}} g_{O(j)},$$

where for $j \geq 1$

$$\text{logit} Q(\epsilon)_{O(j)}(1) = \text{logit} Q_{O(j)}(1) + \epsilon_j H_{O(j)}(Q)$$

is a logistic regression model using the logit of $P_{O(j)}(1 \mid Pa(O(j)))$ as offset, and $H_{O(j)}(Q)$ as d-dimensional covariate of the same dimension as the target parameter ψ_0. Recall that $H_{O(j)}(Q, g) = H_{O(j)}(Q) H_{O(j)}(g)$ and that we now only use the $H_{O(j)}(Q)$-factor. Here ϵ_j is a subvector of ϵ. In addition, the fluctuation model $Q(\epsilon)_{O(0)}$ for the distribution of the baseline covariates $O(0)$ is chosen to have a score of $D^*_{O(0)}(Q)$ with respect to ϵ_0.

Let $L_g(Q) = -\log Q_{O(0)} + \sum_{j \in \mathcal{I}^c, j \geq 1} \{\log Q_{O(j)}\} H_{O(j)}(g)$ be the *weighted* log-likelihood loss function. Since $H_{O(j)}(g)$ is only a function of O through $Pa(O(j))$, it follows that $\arg\min_Q P_0 L_g(Q) = Q_0$ and is thus always a valid loss function (even if g is misspecified). Indeed, the score $\frac{d}{d\epsilon} L_g(Q(\epsilon))$ of $Q(\epsilon)$ at $\epsilon = 0$ equals or spans (if ϵ is multivariate) $D^*(Q, g)$, and the score of $Q_{O(j)}(\epsilon_j)$ at $\epsilon_j = 0$ equals $D^*_{O(j)}(P)$. Thus, $\langle D^*(Q, g) \rangle \subset \langle < \frac{d}{d\epsilon} L_g(Q(\epsilon))\big|_{\epsilon=0} \rangle$.

Note that the weighted maximum likelihood estimator $\epsilon_{jn} = \arg\min_{\epsilon_j} P_n L_g(Q(\epsilon))$ of ϵ_j for a given initial $P = Qg$ can be determined with univariate logistic regression software regressing the binary $O(j)$ on the clever covariate $H^*_{O(j)}(Q)$, using the initial as offset, and using as weights $H^*_{O(j)}(g)(O_i)$, $i = 1, \ldots, n$. If one uses a common ϵ, i.e., $\epsilon_j = \epsilon$ for $j > 0$, then one can fit this single ϵ by regressing the binary outcome $O(j)$ on the clever covariate $H^*_{O(j)}(Q)$ based on a *pooled* data set, using corresponding weights $H^*_{O(j)}(g)(O_i)$, so that all j-specific logistic regressions with common parameter ϵ are fit in one run.

Consider an initial estimator $P^0_n = Q^0_n g^0_n$ of P_0, where $Q^0_{n,O(0)}$ is the empirical distribution of the baseline covariates $O_i(0)$, $i = 1, \ldots, n$. We will use a separate ϵ_0 for the fluctuation of $Q^0_{n,O(0)}$, and it will always equal 0, so that this empirical distribution will not be updated. Given the loss function $L(P) = -\log Q(P)$, we determine

$$\epsilon^1_n = \arg\min_\epsilon P_n L_{g^0_n}(P^0_n(\epsilon)).$$

This results in the first-step TMLE $P^1_n = P^0_n(\epsilon^1_n)$. This updating process $P^k_n = P^{k-1}_n(\epsilon^k_n)$, $k = 1, \ldots, K$, is iterated to convergence defined by $\epsilon^k_n \approx 0$. The final update

is the TMLE of P_0 and is denoted with $P_n^* = Q_n^* g_n^0$. We note that g_n^0 is not updated in this process due to the fluctuation working model only allowing fluctuations of Q_n^0. The TMLE of ψ_0 is now defined as the substitution estimator $\Psi(P_n^*) = \Psi(Q_n^*)$.

One may use a separate ϵ_j for each factor $Q_{n,O(j)}^0$, $j \geq 1$, and determine the weighted maximum likelihood estimators of these fluctuation parameters recursively, starting with the last factor and working backward to the first factor, always using the most recent update. In principle, one would start over at the last factor after having finished the update of the first factor and iterate this updating process until convergence. However, as above for TMLE I, it follows that the TMLE requires only one update per factor and thereby converges in one step (representing one round from the last factor to the first factor) and exists in analytic form. We refer the interested reader to the forthcoming Stitelman and van der Laan (2011a) for implementation of TMLE II.

A.9 Efficient Influence Curve Collaborative Double Robustness

Summary. By definition, the canonical gradient is orthogonal to the scores generated by parametric submodels through the data-generating distribution that do not fluctuate the target parameter of interest. That is, the canonical gradient is orthogonal to the nuisance tangent space. This property implies that the canonical gradient at certain misspecified data-generating distributions will still have a mean of zero under the true data-generating distribution, which is called the robustness of the efficient influence curve with respect to misspecification of nuisance parameters. Robustness of the efficient influence curve translates into robustness of estimators that utilize the efficient influence curve, such as estimating-equation-based estimators and TMLEs. We prove a so-called collaborative double robustness property of the efficient influence curve, which is utilized in the C-TMLE.

Let us denote the intervention nodes by $A(j)$, and the nodes in between two subsequent intervention nodes $A(j-1)$ and $A(j)$ by $L(j)$ so that $O(0), \ldots, O(J)$ is represented as $L(0), A(0), \ldots, A(K), L(K+1)$. Our last representation (A.5) of the canonical gradient is then given by

$$D^*(P_0)(O) = E_{Q_0}\left(\textstyle\sum_{\bar{a}} h \mid L(0)\right) - \Psi(Q_0) + \sum_{j=1}^{K+1} \frac{1}{g_{0,\bar{A}(j-1)}} \left\{E_{Q_0}\left(\textstyle\sum_{\bar{a}(j,K)} h \mid L(j), Pa(L(j))\right) - E_{Q_0}\left(\textstyle\sum_{\bar{a}(j,K)} h \mid Pa(L(j))\right)\right\}, \quad \text{(A.6)}$$

where $\bar{a}(j,K) = (a(j), \ldots, a(K))$. We will use this last representation (A.6) of the efficient influence curve to explicitly prove and demonstrate its collaborative double robustness property. Collaborative double robustness of $D^*(P_0) = D^*(Q_0, g_0)$ can be formulated as follows. For each $Q \in \mathcal{Q}$, and a corresponding specified set $\mathcal{G}(Q, P_0) \subset \mathcal{G}$, we have that

$$g \to P_0 D^*(Q, g)$$

is constant in $g \in \mathcal{G}(Q, P_0)$, and it equals zero if Q is such that $\Psi(Q) = \Psi(Q_0)$. Firstly, we note that $P_0 D^*(Q, g_0) = \psi_0 - \Psi(Q)$, which also follows from the proof below. This shows that $\mathcal{G}(Q, P_0)$ contains, at least, g_0. However, we wish to determine a richer set $\mathcal{G}(Q, P_0)$ of conditional distributions for which $P_0 D^*(Q, g) = \psi_0 - \Psi(Q)$. We will prove the following result.

Result 1 *We have $O = (A, L_A) \sim P_0 = Q_0 g_0$ is a missing-data structure on full-data $X = (L_a : a)$, where $g_0(A \mid X) = \prod_{j=0}^{K} g_{A(j),0}(A(j) \mid \bar{A}(j-1), X)$ satisfies SRA. We define a reduced collection of counterfactual random variables $X^{r*} = X^{r*}_{Q-Q_0} = (\bar{X}^{r*}_{Q-Q_0}(j) : j)$ (i.e., X^{r*} is a function of X), where*

$$\bar{X}^{r*}_{Q-Q_0}(j) = \left\{ E_{Q-Q_0}\left(\textstyle\sum_{\bar{a}(j,K)} h \mid \bar{L}_{\bar{a}(j-1)}(j), \bar{A}(j-1) = \bar{a}(j-1) \right) \right.$$
$$\left. - E_{Q-Q_0}\left(\textstyle\sum_{\bar{a}(j,K)} h \mid \bar{L}_{\bar{a}(j-2)}(j-1), \bar{A}(j-1) = \bar{a}(j-1) \right) : \bar{a}(j-1) \right\}.$$

Let $\mathcal{G}(Q, P_0)$ be all true (i.e., under P_0) conditional distributions of $\bar{A}$, given a reduction X^r that implies X^{r*}; such true conditional probability distributions are related to g_0 by the relation $g_0(\bar{a} \mid X^r) = E_0(g_0(\bar{a} \mid X) \mid X^r)$. Then, for any Q and $g \in \mathcal{G}(Q, P_0)$, we have

$$P_0 D^*(Q, g) = \psi_0 - \Psi(Q).$$

Proof. Firstly, we note that for any g, $P_0 D^*(Q_0, g) = 0$, by simply conditioning on $Pa(L(j))$ for each j-specific term. Therefore, it suffices to determine the set of g for which $P_0\{D^*(Q, g) - D^*(Q_0, g)\} - \{\psi_0 - \Psi(Q)\} = 0$. The left-hand side equals

$$P_0 E_{Q-Q_0}\left(\textstyle\sum_{\bar{a}} h \mid L(0) \right)$$
$$+ \textstyle\sum_{j=1}^{K+1} P_0 \frac{1}{g_{\bar{A}(j-1)}} \left\{ E_{Q-Q_0}\left(\textstyle\sum_{\bar{a}(j,K)} h \mid L(j), Pa(L(j)) \right) - E_{Q-Q_0}\left(\textstyle\sum_{\bar{a}(j,K)} h \mid Pa(L(j)) \right) \right\}.$$

We now utilize the missing-data-structure representation of the observed data $O = (A, L_A)$ as a function of A and the collection of counterfactuals $X = (L_a : a)$, so that, for example, $L(j), Pa(L(j))$ is a function of $(\bar{L}_a(j) : a)$ and $\bar{A}(j-1)$. Thus, the conditional expectations with respect to $Q - Q_0$ are functions of $(\bar{L}_a(j) : a)$ and $\bar{A}(j-1)$. Specifically, at $\bar{A}(j-1) = \bar{a}(j-1)$, the first term of the j-th term equals $E_{Q-Q_0}(\sum_{\bar{a}(j,K)} h \mid \bar{L}_{\bar{a}(j-1)}(j), \bar{A}(j-1) = \bar{a}(j-1))$, and is thus indeed a function of the counterfactual $\bar{L}_{\bar{a}(j-1)}(j)$. Suppose that $g \in \mathcal{G}(Q, P_0)$. Consider now the expectation under P_0 for the jth term, and first take the conditional expectation of $\bar{A}(j-1)$ under g, thereby conditioning on a reduction X^r of X that is rich enough to make the E_{Q-Q_0} terms fixed. This yields

$$P_0 E_{Q-Q_0}\left(\textstyle\sum_{\bar{a}} h \mid L(0) \right) + \textstyle\sum_{j=1}^{K+1} P_0 \sum_{\bar{a}(j-1)} E_{Q-Q_0}\left(\textstyle\sum_{\bar{a}(j,K)} h \mid \bar{L}_a(j), \bar{a}(j-1) \right)$$
$$- \textstyle\sum_{j=1}^{K+1} P_0 \sum_{\bar{a}(j-1)} E_{Q-Q_0}\left(\textstyle\sum_{\bar{a}(j,K)} h \mid \bar{L}_a(j-1), \bar{a}(j-1) \right).$$

Note that this no longer depends on g. The jth term of the first sum and the $j+1$-th term of the second sum gives the following difference:

$$P_0 \sum_{\bar{a}(j-1)} \left\{ E_{Q-Q_0}\left(\sum_{\bar{a}(j,K)} h \mid \bar{L}_a(j), \bar{a}(j-1)\right) - \sum_{a(j)} E_{Q-Q_0}\left(\sum_{\bar{a}(j+1,K)} h \mid \bar{L}_a(j), \bar{a}(j)\right)\right\}.$$

By SRA we have that

$$E_{Q-Q_0}\left(\sum_{\bar{a}(j+1,K)} h \mid \bar{L}_a(j), \bar{a}(j)\right) = E_{Q-Q_0}\left(\sum_{\bar{a}(j+1,K)} h \mid \bar{L}_a(j), \bar{a}(j-1)\right).$$

We can now bring in the sum over $a(j)$:

$$\sum_{a(j)} E_{Q-Q_0}\left(\sum_{\bar{a}(j+1,K)} h \mid \bar{L}_a(j), \bar{a}(j-1)\right) = E_{Q-Q_0}\left(\sum_{\bar{a}(j,K)} h \mid \bar{L}_a(j), \bar{a}(j-1)\right).$$

This proves that the first term of the jth term and the second term of the $j+1$-th term cancel out. In particular, the very first term $P_0 E_{Q-Q_0}(\sum_{\bar{a}} h \mid L(0))$ cancels out with the second term for $j = 1$. This shows that we are left with a single term, namely, the $j = K+1$-th term of the first sum: $P_0 \sum_{\bar{a}(K)} E_{Q-Q_0}(h \mid \bar{L}_a(K+1), \bar{a}(K))$. However, the conditioning event (both under Q and Q_0) implies the value of h so that this conditional expectation E_{Q-Q_0} equals $h - h = 0$. This proves the desired collaborative double robustness. □

A.10 Example: TMLE with the Outcome Subject to Missingness

Suppose $O = (W, A, \Delta, Y^* = \Delta Y) \sim P_0$. The model for the probability distribution of (W, A, Y) is nonparametric so that the model for P_0 is nonparametric beyond the special structure that Y^* equals 0 when $\Delta = 0$. We have that the likelihood of O under P factorizes as $P = P_W P_A P_\Delta P_{Y^*}^\Delta$, using our notation for conditional distributions, where each of these conditional distributions is unspecified, which defines the statistical model $\mathcal{M}$ for P_0. Let $\Psi(P) = E_P[E_P(Y^* \mid W, A = 1, \Delta = 1) - E_P(Y^* \mid W, A = 0, \Delta = 1)]$ be the target parameter of P defined on this model $\mathcal{M}$. We wish to determine the efficient influence curve $D^*(P)$ of $\Psi : \mathcal{M} \to \mathbb{R}$ at a P, and subsequently a TMLE of ψ_0. We note that $P = Q \times g$, with g being the conditional distribution of (A, Δ), given W, $Q = Q_W Q_{Y^*}$, is a product of the other two factors of P, and we also note that $\Psi(P) = \Psi(Q)$.

As the initial gradient of the pathwise derivative of Ψ at $P = Q \times g$ in the model with g known, we can choose $D_{IPCW}(P) = H(g)Y - \Psi(P)$, where $H(g) = I(A = 1, \Delta = 1)/g(A, \Delta \mid W) - I(A = 0, \Delta = 1)/g(A, \Delta \mid W)$. This is a gradient in the model with g known, but its projection onto the tangent space of Q will yield the efficient influence curve in our model. One way to verify that D_{IPCW} is a valid IPCW gradient is to use a missing-data-structure representation of $O = (W, A, \Delta, \Delta Y_A)$ on full-data $X = (W, Y_0, Y_1)$, use the theory for CAR censored-data models as presented in van der Laan and Robins (2003) for determining IPCW gradients, and apply the invariance principle presented in Appendix A.7. That is, (1) assume CAR, $P(A, \Delta \mid X) = P(A, \Delta \mid W)$, (2) note that $\Psi(P) = EY_1 - EY_0$ is a parameter of full-data distribution, (3) note that the gradient of the full-data parameter $EY_1 - EY_0$ in the full-data model for X is given by $D^F(X) = (Y_1 - Y_0 - \psi)$, and (4) show that $E(D_{IPCW} \mid X) = D^F(X)$. We refer to van der Laan and Robins (2003) for detailed

theory on determining gradients for censored-data models in terms of the gradients of the underlying full-data model.

The efficient influence curve $D^*(P)$ equals the projection of D_{IPCW} onto the tangent space of Q. The projection onto the tangent space of Q_W is given by $E(D_{IPCW} \mid W)$, and the projection onto the tangent space of $Q_{Y^*|W,A,\Delta}$ is given by $\Delta\{D_{IPCW} - E(D_{IPCW} \mid W, A, \Delta = 1)\}$. Since the sum of these two projections yields the efficient influence curve $D^*(P)$, we have

$$D^*(P) = H(g)(Y^* - \bar{Q}(W, A)) + \bar{Q}(W, 1) - \bar{Q}(W, 0) - \Psi(Q),$$

where $\bar{Q}(W, a) = E_P(Y^* \mid W, A = a, \Delta = 1)$.

The TMLE. If Y, and thereby Y^*, is binary, then we use the fluctuation working model $\text{logit}\bar{Q}(\epsilon) = \text{logit}\bar{Q} + \epsilon H(g)$ and use as loss function for $\bar{Q}$ the log-likelihood for a binary distribution given by $L(\bar{Q}) = \bar{Q}^{Y^*\Delta}(1 - \bar{Q})^{(1-Y^*)\Delta}$. We can also propose a fluctuation working model for Q_W with score $D_W = \Pi(D^* \mid T_W)$, but since we will use as initial estimator of $Q_{W,0}$ the empirical distribution, the maximum likelihood estimator of the fluctuation parameter will be zero, so that no updates of $Q_{W,n}$ will occur. Let $\bar{Q}_n^0$ be an initial estimator of $\bar{Q}_0$. One now computes $\epsilon_n = \arg\max_\epsilon P_n L(\bar{Q}_n^0(\epsilon))$, and one defines the TMLE update as $\bar{Q}_n^1 = \bar{Q}_n^0(\epsilon_n)$. Further iteration does not result in further updates, so that the TMLE Q_n^* is defined as $Q_n^* = (Q_{W,n}, \bar{Q}_n^1)$. The TMLE of ψ_0 is thus given by $\Psi(Q_n^*)$. If Y is continuous with values in $(0, 1)$, one can use the same loss function $L(\bar{Q})$ and fluctuation function for the conditional mean $\bar{Q}$. If Y is bounded in (a, b), then one can transform the outcome into $Y^* = (Y - a)/(b - a) \in (0, 1)$ and apply this same TMLE. For continuous Y, one can also use the squared error loss function and the linear fluctuation function, but such a fluctuation function is not guaranteed to respect known bounds on Y and is thus not generally recommended.

A.11 Example: TMLE of Causal Effect in a Two-Stage RCT

Let us denote the observed data structure on a randomly sampled patient from the target population with $O = (L(0), A(0), L(1), A(1), Y = L(2)) \sim P_0$. Let $Pa(L(j)) = (\bar{A}(j-1), \bar{L}(j-1))$ and $Pa(A(j)) = (\bar{L}(j), \bar{A}(j-1))$, where we make the convention that for $j = 0$, $\bar{A}(j-1)$ and $\bar{L}(j-1)$ are empty. The likelihood of O can be factorized as $P = P_{L(0)} \prod_{l=1}^{L} P_{L(1,l)} P_Y \prod_{j=0}^{1} P_{A(j)}$, where the first factors will be denoted by $Q_{L(0)}$, $Q_{L(1,l)}$, Q_Y, and the latter two factors denote the treatment mechanism and are denoted by $g_{A(j)}$, $j = 0, 1$.

Suppose our parameter of interest is the treatment-specific mean EY_d for a certain treatment rule d that assigns treatment $d_0(L(0))$ at time 0 and treatment $d_1(\bar{L}(1), A(0))$ at time 1. For example, $d_0(L(0)) = 1$ is a static treatment assignment, $L(1)$ is binary, and $d_1(\bar{L}(1), A(0)) = I(L(1) = 1)\times 1 + I(L(1) = 0)\times 0$ assigns treatment 1 if the patients responds well to the first line treatment (i.e., $L(1) = 1$) and treatment 0 if the patient does not respond well to the first line treatment. We

note that any treatment rule can be viewed as a function of $\bar{L} = (L(0), L(1))$ only, and therefore we will use the shorter notation $d(\bar{L}) = (d_0(L(0)), d_1(\bar{L}))$ for the two rules at times 0 and 1.

Note that $EY_d = \Psi(Q)$ for a well-defined mapping Ψ. Specifically, we have $\Psi(Q) = E_{P_d}Y$, where the postintervention distribution P_d of $(L(0), L(1), L(2))$ is defined by the g-computation formula: $P_d(\bar{L}) = \prod_{j=0}^{2} Q_{L(j),d}(\bar{L}(j))$, where, for notational convenience, we used the notation $Q_{L(j),d}(\bar{L}(j)) = Q_{L(j)}(L(j) \mid \bar{L}(j-1), \bar{A}(j-1) = d(\bar{L}(j-1)))$. From this analytic expression it also follows that, even if Y is continuous, $\Psi(Q)$ only depends on the conditional distribution of Y through its mean. Using the techniques given above we obtain the following representation of the efficient influence curve.

Theorem A.2. *The efficient influence curve for $\psi = EY_d$ at the distribution $P = Qg$ of O can be represented as $D^* = \Pi(D_{IPCW} \mid T_Q)$, where $D_{IPWC}(O) = \frac{I(\bar{A}=d(\bar{L}))}{\prod_{j=0}^{1} g_{A(j)}} Y - \psi$, T_Q is the tangent space of Q in the nonparametric model, and Π denotes the projection operator onto T_Q in the Hilbert space $L_0^2(P)$ of square P-integrable functions of O, endowed with inner product $\langle h_1, h_2 \rangle = E_P h_1 h_2(O)$. We have that $T_Q = \sum_{j=0}^{2} T_{Q_{L(j)}}$ is the orthogonal sum of the tangent spaces $T_{Q_{L(j)}}$ of the $Q_{L(j)}$-factors, which consists of functions of $(L(j), Pa(L(j)))$ with conditional mean zero, given the parents $Pa(L(j))$ of $L(j)$, $j = 0, 1, 2$. Recall that we also denote $L(2)$ by Y. Let $D_{L(j)} = \Pi(D^* \mid T_{Q_{L(j)}})$, $j = 0, 1, 2$. We have $D_{L(1)} = \sum_{l=1}^{L} D_{L(1),l}$, where $D_{L(1),l} = \Pi(D \mid T_{Q_{L(1,l)}})$, and*

$$
\begin{aligned}
D_{L(0)}(O) &= E(Y_d \mid L(0)) - \psi, \\
D_{L(1,l)}(O)) &= \frac{I\left(A(0) = d_0(L(0))\right)}{g_{A(0)}} \left\{C_{L(1,l)}(1) - C_{L(1,l)}(0)\right\} \left\{L(1,l) - Q_{L(1,l)}(1)\right\}, \\
D_{L(2)}(O) &= \frac{I\left(\bar{A} = d(\bar{L})\right)}{\prod_{j=0}^{1} g_{A(j)}} \left\{L(2) - E(L(2) \mid \bar{L}(1), \bar{A}(1))\right\},
\end{aligned}
$$

where, for $\delta \in \{0, 1\}$, $C_{L(1,l)}(\delta) = E(Y_d \mid Pa(L(1,l)), L(1,l) = \delta)$. We also note that: $E(Y_d \mid L(0), A(0) = d_0(L(0)), L(1)) = E(Y \mid \bar{L}(1), \bar{A} = d(\bar{L}))$.

The TMLE. Consider now an initial estimator $Q_{L(j),n}$ of each $Q_{L(j)}$, $j = 0, 1, 2$. We will estimate the first marginal probability distribution $Q_{L(0)}$ of $L(0)$ with the empirical distribution of $L_i(0)$, $i = 1, \ldots, n$. We use the log-likelihood loss function $L(Q_{L(0)}) = -\log Q_{L(0)}$ and consider a submodel $Q_{L(0)}(\epsilon_0)$ with score $D_{L(0)}(Q)$ defined above.

We can estimate the conditional distributions of the binary $L(1, l)$ with loss-based learning based on the loss function $L(Q_{L(1)}) = -\sum_l \log Q_{L(1,l)}$. For example, one could use logistic regression machine learning algorithms, $l = 1, \ldots, L$, where one could also smooth in l. Similarly, we can estimate the conditional mean of $Y = L(2)$ with loss-based learning using the log-likelihood or squared error loss function. We will now define fluctuations of this initial estimator $Q_{L(1),n}$ and $Q_{L(2),n}$. Firstly, let

$$
\text{logit}\bar{Q}_{L(1,l),n}(\epsilon_1) = \text{logit}\bar{Q}_{L(1,l),n} + \epsilon_1 H^*_{L(1,l)}(Q_n, g_n)
$$

be the fluctuation function of the conditional probability $\bar{Q}_{L(1,l),n} = Q_{L(1,l),n}(1)$ of $L(1,l) = 1$ with fluctuation parameter ϵ_1, where we added the covariate $H^*_{L(1,l)}(Q,g) = \{I(A(0) = d(L(0)))/g_{A(0)}\}(C_{L(1,l)}(1) - C_{L(1,l)}(0))$ defined above in Theorem A.2. This defines a fluctuation working model $Q_{L(1),n}(\epsilon_1)$, and we can use the log-likelihood $-\log Q_{L(1)}$ as loss function. In the special case where $L(1)$ is itself already a binary variable, we have

$$H^*_{L(1)}(L(0),A(0)) = \frac{I(A(0)=d_0(L(0)))}{g_{A(0)}(d_0(L(0))|L(0))}\{C_{L(1)}(Q)(1) - C_{L(1)}(Q)(0)\},$$

where $C_{L(1)}(Q)(\delta) = E_Q(Y_d \mid L(0), A(0), L(1) = \delta)$. We refer to these covariate choices as *clever* covariates, since they represent a covariate choice that identifies a least favorable fluctuation model, thereby providing the desired targeted bias reduction. Similarly, if $Y = L(2)$ is binary, then let $\text{logit}\bar{Q}_{L(2)n}(\epsilon_2) = \text{logit}\bar{Q}_{L(2)n} + \epsilon_2 H^*_{L(2)}(Q_n, g_n)$, where the added clever covariate $H^*_{L(2)}(Q,g)(\bar{L}(1),\bar{A}(1)) = I(\bar{A} = d(\bar{L}))/\prod_{j=0}^{1} g_{A(j)}$. If Y is continuous with values in $(0,1)$, then we can use the same "log-likelihood" loss function $L(\bar{Q}_Y) = -\{Y \log \bar{Q}_Y + (1-Y)\log(1-\bar{Q}_Y)\}$, and fluctuation working model as we use for binary Y, and this yields then also a TMLE for Y bounded in (a,b). As a side remark, one may also select the squared error loss function for $\bar{Q}_Y$ and linear fluctuation working model $\bar{Q}_{L(2),n} + \epsilon H_{L(2)}(Q_n, g_n)$. We note that the above fluctuation function and use of the loss function $L(Q) = L(Q_{L(0)}) - \sum_l \log Q_{L(1,l)} + L(\bar{Q}_Y)$ indeed satisfies that the score of $\epsilon = (\epsilon_0, \epsilon_1, \epsilon_2)$ at $\epsilon = 0$ spans the efficient influence curve $D^*(Q_n, g_n)$, as presented in Theorem A.2 above.

Let $\epsilon_n = \arg\max_\epsilon \prod_{j=1}^{2}\prod_{i=1}^{n} Q_{L(j),n}(\epsilon)(O_i)$ be the maximum likelihood estimator of ϵ according to the working fluctuation model. If one uses separate ϵ for different factors of Q, then one could also obtain a separate maximum likelihood estimator of ϵ_j for each factor $j = 1,2$, or even an $\epsilon_{1,l}$ for each factor $Q_{L(1,l)}$ indexed by l. This process is now iterated to convergence, which defines the TMLE (Q^*_n, g_n), starting at initial estimator (Q_n, g_n). Note that this does not involve updating of g_n. The TMLE of ψ is now given by $\Psi(Q^*_n)$.

If Y is binary or continuous in $(0,1)$, then a single/common ϵ_n defined above requires applying a single logistic regression applied to a repeated measures data set with one line of data for each of the factors of the likelihood, creating a clever covariate column that alternates the clever covariates $H_{L(1,l)}$ and $H_{L(2)}$, and uses the corresponding offsets. Thus, in both cases (separate or common ϵ), the update step can be carried out with a simple univariate logistic regression maximum likelihood estimator. Computing a MLE of a common ϵ in the case where we use the linear fluctuation working model and squared error loss function for Q_Y requires some programming.

A one-step closed-form TMLE. For the sake of illustration, suppose $L(1)$ is itself binary. If one uses a separate $\epsilon_{L(j)}$ for $j = 0,1,2$, first carry out the TMLE update for $Q_{L(2),n}$ and use this updated $Q^*_{L(2),n}$ in the clever covariate required to compute the targeted update of $Q_{L(1),n}$. Then we obtain a TMLE algorithm that converges in these two simple updating steps, representing a single-step update of Q^*_n and

TMLE $\Psi(Q_n^*)$. Similarly, this particular one-step TMLE involving iterative updating (starting with the last factor of the likelihood and ending with the update of the first factor) generalizes to general $L(1)$ and general longitudinal data structures (van der Laan 2010a), and was presented above.

A.12 Example: TMLE with Right-Censored Survival Time

Let $O = (W, A, dN(t), dA_c(t), t = 1, \ldots, \tau) \sim P_0$, where $dN(t) = I(\tilde{T} = t, \Delta = 1)$ and $dA_c(t) = I(\tilde{T} = t, \Delta = 0)$, are indicators of an observed failure and observed censoring event at time t, respectively. The likelihood of O under P factorizes as $P = Q_W g_A \prod_t Q_{dN(t)} \prod_t g_{dA_c(t)}$, using our short-hand notation, where $Pa(dN(t)) = (W, A, \bar{N}(t-1), \bar{A}_c(t-1))$, $Pa(A) = W$, and $Pa(dA_c(t)) = (W, A, \bar{A}_c(t-1), \bar{N}(t))$. The model for P_0 is nonparametric on the Q_0-factor but may incorporate assumptions about $g_0 = g_{0,A} \prod_t g_{0,dA_c(t)}$. The efficient influence curve of a parameter of Q is the same in any such a model. Let $\Psi(P) = E_P[S(t_0 \mid A = 1, W) - S(t_0 \mid A = 0, W)]$, where $S(t_0 \mid A = a, W) = \prod_{t=0}^{t_0}(1 - \bar{Q}_{dN(t)}(t \mid W, A = a))$, and $\bar{Q}_{dN(t)}(t \mid W, A) = E_P(dN(t) \mid \tilde{T} \geq t, W, A)$ would be equal to the conditional hazard of an underlying time-to-event T under a CAR-model (see below). We wish to determine the efficient influence curve $D^*(P)$ of this target parameter at P and define a TMLE.

As initial gradient in the model in which g is known, we can choose

$$D_{IPCW} = \frac{I(A=1,\bar{A}_c(t_0)=\bar{0}(t_0))-I(A=0,\bar{A}_c(t_0)=\bar{0}(t_0))}{g_A \prod_{t \leq t_0} g_{dA_c(t)}} I(\tilde{T} > t_0) - \Psi(Q).$$

One way to show that this is indeed a gradient in the model with g known is (based on van der Laan and Robins 2003) (1) to represent O as a missing-data structure on full data structure $X = (W, T_0, T_1)$ with censoring process $(A, I(C = t) : t = 1, \ldots, \tau)$, so that $T = T_A$, $dN(t) = I(T_A = t, C \geq t)$, and $dA_c(t) = I(C = t, T_A > t)$; (2) assuming CAR on the conditional distribution of (A, A_c), given X, so that $P(A \mid X) = P(A \mid W)$ and $P(dA_c(t) \mid X, \bar{A}_c(t-1), A) = P(dA_c(t) \mid \bar{A}_c(t-1), \bar{N}(t), W, A)$; (3) noting that $\Psi(P) = P(T_1 > t_0) - P(T_0 > t_0)$ and that the gradient of this full-data parameter in the full data model equals $D^F(X) = I(T_1 > t_0) - I(T_0 > t_0) - \Psi(P)$; and (4) showing that $E(D_{IPCW} \mid X) = D^F(X)$. Such a CAR censored-data representation of the observed data model provides no restrictions on the statistical model, so that its only role is to provide a working model for carrying out calculations, such as the calculation of an efficient influence curve (Appendix A.7).

The efficient influence curve $D^*(P)$ equals the projection of $D_{IPCW}(P)$ onto the tangent space of $Q = Q(P)$. The projection onto the tangent space $T_W(P)$ of Q_W equals $E_P(D_{IPCW} \mid W)$, which equals $S(t_0 \mid A = 1, W) - S(t_0 \mid A = 0, W) - \Psi(Q)$. The projection onto the tangent space $T_{dN(t)}(P)$ of $Q_{dN(t)}$ is given by $H^*_{dN(t)}(dN(t) - \bar{Q}_{dN(t)}(t \mid W, A))$, where

$$H^*_{dN(t)} = E_P(D_{IPCW} \mid dN(t) = 1, Pa(dN(t))) - E_P(D_{IPCW} \mid dN(t) = 0, Pa(dN(t))).$$

Since $Pa(dN(t)) = (W, A, \bar{N}(t-1), \bar{A}_c(t-1))$ and the projection onto $T_{dN(t)}$ equals zero if $\tilde{T} \leq t-1$, it follows that we can condition on $\tilde{T} \geq t$ in these two conditional expectations. For $t \leq t_0$ we have $E_P(D_{IPCW} \mid W, A, dN(t) = 1, \tilde{T} \geq t) = 0$, since $dN(t) = 1$ implies $\tilde{T} \leq t_0$ so that $D_{IPCW} = 0$. For $t > t_0$, this same conditional expectation reduces to $(I(A=1) - I(A=0)) / \left(g_A \prod_{s \leq t_0} g_{dA_c(s)}(0)\right)$. Regarding the first conditional expectation in the expression for $H_{dN(t)}$, we conclude

$$E_P(D_{IPCW} \mid W, A, dN(t) = 1, \tilde{T} \geq t) = I(t > t_0) \frac{I(A=1) - I(A=0)}{g_A \prod_{s \leq t_0} g_{dA_c(s)}(0)}.$$

Let us now consider the second conditional expectation in $H_{dN(t)}$. If $t > t_0$, then this term equals the first term we just displayed. For $t \leq t_0$, we obtain

$$E_P(D_{IPCW} \mid W, A, dN(t) = 0, \tilde{T} \geq t)$$
$$= \frac{I(A=1)-I(A=0)}{g_A \prod_{s \leq t-1} g_{dA_c(s)}(0)} E_P\left(\frac{I(\bar{A}_c(t,t_0)=0, \tilde{T} > t_0)}{\prod_{s \in [t,t_0]} g_{dA_c(s)}} \mid W, A, \tilde{T} \geq t, dN(t) = 0 \right).$$

In the CAR censored-data model representation, the latter conditional expectation equals $E_P(I(T > t_0) \mid W, A, T > t)$. Regarding the second term, we conclude:

$$E_P(D_{IPCW} \mid W, A, dN(t) = 0, \tilde{T} \geq t)$$
$$= I(t > t_0) \frac{I(A=1)-I(A=0)}{g_A \prod_{s \leq t_0} g_{dA_c(s)}(0)} + I(t \leq t_0) \frac{I(A=1)-I(A=0)}{g_A \prod_{s \leq t-1} g_{dA_c(s)}(0)} \frac{S(t_0|W,A)}{S(t|W,A)}.$$

Thus, we have shown that

$$H^*_{dN(t)} = -I(t \leq t_0) \frac{I(A=1)-I(A=0)}{g_A \prod_{s \leq t-1} g_{dA_c(s)}(0)} \frac{S(t_0|W,A)}{S(t|W,A)}$$

and that the efficient influence curve can be represented as

$$D^*(P) = \sum_{t=1}^{\tau} H_{dN(t)}(dN(t) - \bar{Q}_{dN(t)}) + S(t_0 \mid A = 1, W) - S(t_0 \mid A = 0, W) - \Psi(Q).$$

The TMLE. In our chapters on TMLE for time-to-event outcomes, we presented the TMLE based on this representation of the efficient influence curve, involving TMLE updates of an estimator of the conditional hazard $\bar{Q}_{dN(t)}$ using a logistic regression working fluctuation model with a time-dependent clever covariate $H^*_{dN(t)}$.

A.13 Example: TMLE of a Causal Effect Among the Treated

Suppose we observe n i.i.d. observations of $O = (W, A, Y) \sim P_0$, W baseline covariates, subsequently assigned binary treatment A, and final outcome Y of interest. Suppose the statistical model is nonparametric and we wish to estimate the following parameter of the data-generating distribution P_0 of $O = (W, A, Y)$:

$$\Psi(P_0) = E_0 \left[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W) \mid A = 0 \right].$$

Another way of representing this parameter is $\Psi(P_0) = -E_0(Y - E(Y \mid A = 1, W) \mid A = 0)$, i.e., among the nontreated in the population, one evaluates the outcome minus the predicted outcome if, contrary to fact, one would have been treated, and one takes the population average of all these differences over all nontreated subjects. Under an SCM, $W = f_W(U_W)$, $A = f_A(W, U_A)$, $Y = f_Y(W, A, U_Y)$, and the randomization assumption stating that U_A is independent of U_Y, one can interpret this parameter as a causal effect among the nontreated $E(Y_1 - Y_0 \mid A = 0)$. Suppose one wants to estimate the effect among the treated, given by

$$\Psi_1(P_0) = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W) \mid A = 1],$$

which under the above-mentioned SCM can be represented as $E(Y_1 - Y_0 \mid A = 1)$. Switching the roles of $A = 1$ and $A = 0$ in the formulas below provides the efficient influence curve and TMLE of $-\Psi_1(P_0)$. We make this explicit below.

Note that a probability distribution P is determined by the marginal distribution P_W of W, the conditional distribution $P_{A|W}$ of A, given W, and the conditional distribution $P_{Y|A,W}$ of Y, given A, W. The parameter $\Psi(P)$ depends on P through both $P_W, P_{Y|A,W}$ as well as the treatment mechanism $P_{A|W}$. We will denote the treatment mechanism by $g = g(P)$ and the other two factors of the likelihood by Q_W and $Q_{Y|A,W}$. We will use the notation $\bar{Q}(A, W) = E_P(Y \mid A, W)$ and $\bar{Q}_0$ for this conditional mean of Y under P_0.

Efficient influence curve of target parameter. Firstly, consider the parameter $P \to \Psi(P)(1) = E_P(E_P(Y \mid A = 1, W) \mid A = 0)$. Using the functional delta method technique presented in Appendix A.3, it follows that the efficient influence curve of this parameter at P is given by

$$D_1^*(P) = \frac{I(A=1)}{P(A=0)} \frac{g(0|W)}{g(1|W)}(Y - \bar{Q}(1, W)) + \frac{I(A=0)}{P(A=0)}(\bar{Q}(1, W) - \Psi(P)(1)).$$

Similarly, the efficient influence curve of $\Psi(P)(0) = E_P(E_P(Y \mid A = 0, W) \mid A = 0)$ at P is given by

$$D_0^*(P) = \frac{I(A=0)}{P(A=0)}(Y - \bar{Q}(0, W)) + \frac{I(A=0)}{P(A=0)}(\bar{Q}(0, W) - \Psi(P)(0)).$$

Thus the efficient influence curve of $\Psi(P) = \Psi(P)(1) - \Psi(P)(0)$ is given by

$$D^*(P) = \left\{ \frac{I(A=1)}{P(A=0)} \frac{g(0|W)}{g(1|W)} - \frac{I(A=0)}{P(A=0)} \right\}(Y - \bar{Q}(A, W)) + \frac{I(A=0)}{P(A=0)} \left\{ \bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(P) \right\}.$$

The efficient influence curve of $\Psi(P) = E_P(E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W) \mid A = 1)$ is obtained by changing the roles of $A = 1$ and $A = 0$, and assigning a minus sign, giving

$$D^*(P) = \left\{ \frac{I(A=1)}{P(A=1)} - \frac{I(A=0)}{P(A=1)} \frac{g(1|W)}{g(0|W)} \right\}(Y - \bar{Q}(A, W)) + \frac{I(A=1)}{P(A=1)} \left\{ \bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(P) \right\}.$$

Collaborative double robustness of efficient influence curve. This efficient influence curve of $\Psi(P)$ can be represented as an estimating function $D^*(Q, g, \psi)$, where we suppress the dependence on the scalar $P(A = 0)$ and use notation $Q = (Q_W, \bar{Q})$.

We note that this estimating function is double robust in the sense that it is an unbiased estimating function for ψ_0, if either Q is correctly specified or g is correctly specified. Formally, this is stated as

$$P_0 D^*(Q, g, \psi_0) = 0 \text{ if } Q = Q_0 \text{ or } g = g_0,$$

and $g(1 \mid W) > 0$, a.e. This double robustness result can be explicitly verified.

In fact, we can establish a stronger collaborative double robustness, defined as follows. Let $W(Q)$ be a subset/reduction of W so that conditioning on $W(Q)$ also fixes $(\bar{Q} - \bar{Q}_0)(a, W)$ for $a \in \{0, 1\}$. Then, for all Q and corresponding $g_0(Q) = P(A = \cdot \mid W(Q))$ for such a $W(Q) \subset W$, we have

$$P_0 D^*(Q, g_0(Q), \psi_0) = 0.$$

Note that this implies, in particular, $P_0 D^*(Q_0, g) = 0$ for all g, since, if $Q = Q_0$, then we can select $W(Q)$ as the empty set. Thus, g_0 only needs to adjust for the covariates that still play a role in $\bar{Q} - \bar{Q}_0$. This can also be stated as the following collaborative double robustness of the efficient influence curve $D^*(P) = D^*(Q, g)$. For a given Q, let $\mathcal{G}(Q, P_0)$ be the set of conditional distributions under P_0 of A, given $W(Q)$ as defined above. For each Q, and for each $g \in \mathcal{G}(Q, P_0)$, we have that $P_0 D^*(Q, g) = 0$ implies $\Psi(Q, g) = \psi_0$.

Implications for double robust efficient estimation. One could define a closed-form asymptotically efficient double robust estimator $\psi_{DR,n}$ as the solution of the efficient influence curve estimating equation

$$0 = P_n D^*(Q_n, g_n, \psi),$$

given estimators Q_n of Q_0 and g_n of g_0. We can also compute a collaborative double robust asymptotically efficient TMLE that has various previously presented advantages. In particular, it is guaranteed to be a substitution estimator, and it will only pursue adjustment in g_n that remains helpful after the adjustment carried out by Q_n, thereby resulting in more effective adjustment sets and bias reduction.

A TMLE is a substitution estimator $\Psi(P_n^*)$, where the estimated data-generating distribution P_n^* is such that it solves the efficient influence curve estimating equation

$$0 = P_n D^*(Q(P_n^*), g(P_n^*), \Psi(P_n^*)).$$

As a consequence, the substitution estimator (TMLE) $\Psi(P_n^*)$ is double robust and efficient, and collaborative double robust if one uses the C-TMLE that builds g_n based on the loss function for $\bar{Q}_0$.

The TMLE. Let us now present the TMLE that maps an initial estimator P_n^0 into a targeted fit P_n^*. Suppose Y is binary. Given an initial estimator $\bar{Q}_n^0$ of $\bar{Q}_0$, an initial estimator g_n^0 of g_0, and empirical distribution $Q_{W,n}$ of W, we define the parametric working model for fluctuating the initial estimator: $\text{logit}\bar{Q}_n^0(\epsilon_1) = \text{logit}\, \bar{Q}_n^0 + \epsilon_1 C_1(g_n^0)$,

and $\text{logit}\,(g_n^0(\epsilon_2)(0 \mid W)) = \text{logit}(g_n^0(0 \mid W)) + \epsilon_2 C_2(P_n^0)(W)$, where these two clever covariates are defined as

$$C_1(g) = \left\{ \frac{I(A=1)}{P(A=0)} \frac{g(0 \mid W)}{g(1 \mid W)} - \frac{I(A=0)}{P(A=0)} \right\},$$
$$C_2(P) = \frac{1}{P(A=0)} \left\{ \bar{Q}(P)(1, W) - \bar{Q}(P)(0, W) - \Psi(P) \right\}.$$

Let $Q_{W,n}(\epsilon_0)$ be a parametric working model with score $D_W^* = g(0 \mid W)(\bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(P))$. These three one-dimensional working models represent a parametric working model $\{P_n^0(\epsilon) : \epsilon\}$ for fluctuating P_n^0. We use the log-likelihood loss function $L(P) = -\log P$. We estimate ϵ with maximum likelihood. Note that $\epsilon_{0n} = 0$, ϵ_1 is estimated with standard linear logistic regression fixing $\bar{Q}_n$ as an offset, and ϵ_2 is estimated with standard linear logistic regression fixing $g_n(0 \mid W)$ as offset in the logistic regression model for $P(A = 0 \mid W)$.

This maximum likelihood estimator $\epsilon_n^1 = (\epsilon_{0n}, \epsilon_{1n}, \epsilon_{2n})$ now defines an update $P_n^1 = P_n^0(\epsilon_n^1)$. The targeted maximum likelihood updating is iterated to convergence, and the final P_n^*, identified by a $\bar{Q}_n^*, g_n^*$ and the empirical distribution $Q_{W,n}$, is called the TMLE of the distribution P_0, while $\Psi(P_n^*)$ is called the TMLE of ψ_0. We have that the TMLE $\Psi(P_n^*)$ solves the efficient influence curve estimating equation, as presented above. We can use machine learning/super learning to obtain the initial P_n^0 (i.e., $\bar{Q}_n^0$ and g_n^0).

Since P_n^* solves, in particular,

$$0 = \tfrac{1}{n} \textstyle\sum_{i=1}^n \frac{I(A_i=0)}{P(A=0)} \left\{ \bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i) - \Psi(P_n^*) \right\},$$

it follows that the TMLE $\Psi(P_n^*)$ can also be evaluated as

$$\Psi(P_n^*) = \tfrac{1}{\sum_i I(A_i=0)} \textstyle\sum_i I(A_i = 0) \left\{ \bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i) \right\},$$

i.e., as the empirical mean of $\bar{Q}_n^*(1, W) - \bar{Q}_n^*(0, W)$ among the observations with $A_i = 0$. Apparently, in this evaluation of $\Psi(P_n^*)$, g_n^* can be ignored.

Collaborative TMLE. The collaborative double robustness of the efficient influence curve allows us to also implement the C-TMLE as presented in Chap. 19. A TMLE requires iterative estimation of $\bar{Q}_0$ and g_0, so that other parts of the probability distribution can be ignored. In this case, given the initial estimator Q_n^0 (thus $\bar{Q}_n^0$ and empirical distribution $Q_{W,n}$ of W), one starts with a g_n^0 as an intercept model, and one selects the main term extension g_n^1 of g_n^0 whose TMLE yields the best fit of $\bar{Q}_0$ as measured by the loss function for $\bar{Q}_0$ used by the TMLE. This process is iterated, thereby building a sequence of main term regression fits for g_0 and corresponding TMLEs P_n^k, $k = 1, \ldots, K$. If at a certain k, the loss-function-specific fit of the corresponding TMLE of $\bar{Q}_0$ is not increasing relative to the $k-1$-th TMLE, then we accept the previously selected TMLE, reject the kth TMLE, and proceed as before but now use the $k-1$-th TMLE as initial estimator in the next TMLEs. These subsequent TMLEs will still keep updating the previous g_0 fit by adding main terms.

In this manner, the algorithm generates a sequence of candidate TMLEs indexed by the number of main terms that were included in the g_0 fit. The empirical risk (with respect to the $\bar{Q}_0$ loss function) of these TMLEs decreases with the number of main terms. This number of main terms, and thereby the TMLE, is selected with ($\bar{Q}_0$-) loss-function-specific cross-validation, possibly penalizing the cross-validated risk, as proposed in van der Laan and Gruber (2010) and presented in Chap. 19. The main terms can include propensity score dimension reductions indexed by different adjustment sets, so that the above algorithm is still arbitrarily nonparametric in fitting g_0. We refer the reader to Appendix A.17 for a detailed understanding of the C-TMLE.

A.14 Example: TMLE Based on an Instrumental Variable

Suppose we observe $O = (W, R, A, Y) \sim P_0$. Consider the following SCM: $W = f_W(U_W)$, $R = f(W, U_R)$, $A = f(W, R, U_A)$, $Y = f_Y(W, R, A, U_Y)$. This SCM allows us to define counterfactuals corresponding with setting R and setting simultaneously (R, A), and corresponding postintervention distributions. It is assumed that U_R is independent of U_Y, given W, which means that R is randomized, conditional on W. R plays the role of an instrumental variable that can be used to estimate the causal effect of a treatment A on Y, even if there are unmeasured variables that affect both A and Y (i.e., not captured by W). We consider the following causal parameter of the distribution of counterfactuals corresponding with interventions on R:

$$\Psi_r^F(P_{X,0}) = \frac{E_0 Y(R = r) - EY(R = 0)}{E_0 A(R = r) - E_0 A(R = 0)}. \tag{A.7}$$

By the randomization assumption, this parameter is identifiable from P_0 through the following statistical parameter:

$$\Psi_r(P_0) = \frac{E_0(E_0(Y \mid W, R = r) - E_0(Y \mid W, R = 0))}{E_0(E_0(A \mid W, R = r) - E_0(A \mid W, R = 0))}.$$

We will use the following notation: $\bar{Q}_0(W, r) = E_0(Y \mid W, R = r)$, $\bar{g}_0(W, r) = E_0(A \mid W, R = r)$, $Q_{W,0}(w) = P_0(W = w)$, and $Q_0 = (Q_{W,0}, \bar{Q}_0)$.

Causal interpretation of $\psi_{r,0}$. If the exclusion restriction given by $f_Y(W, R, A, U_Y) = f_Y(W, A, U_Y)$ holds, and $f_Y(W, A, U_Y) = f_Y(W, 0, U_Y) + \beta_0 A$, then it follows that $\beta_0 = \Psi_{r,1}^F(P_{X,0})$. Since the causal interpretation of $\psi_{r,0}$ is constant in r, we can define as estimand a weighted average of the r-specific parameters $\Psi_r(P_0)$, such as $\Psi(P_0) = \sum_{r>0} h(r)\Psi_r(P_0)$, where $\sum_{r>0} h(r) = 1$.

Efficient influence curve. Since $\Psi_r(P_0)$ is a simple function of $E_0 Y(r)$, $E_0 Y(0)$, $E_0 A(r)$, and $E_0 A(0)$, and we know the efficient influence curves of these parameters, the delta method provides us with the efficient influence curve of Ψ_r at P:

$$
\begin{aligned}
D_r^*(P) = & \frac{1}{E\{A(r) - A(0)\}} \frac{I(R = r) - I(R = 0)}{g(R \mid W)} (Y - \bar{Q}(W, R)) \\
& - \frac{E(Y(r) - Y(0))}{E^2(A(r) - A(0))} \frac{I(R = r) - I(R = 0)}{g(R \mid W)} (A - \bar{g}(W, R)) \\
& + \frac{1}{E\{A(r) - A(0)\}} \left\{ \bar{Q}(W, r) - \bar{Q}(W, 0) - E(Y(r) - Y(0)) \right\} \\
& - \frac{E(Y(r) - Y(0))}{E^2(A(r) - A(0))} \left\{ \bar{g}(W, 1) - \bar{g}(W, 0) - E(A(r) - A(0)) \right\}.
\end{aligned}
$$

Again, by the δ-method, the efficient influence curve of Ψ is given by $D^* = \sum_{r>0} h(r) D_r^*$.

Double robustness of r-specific efficient influence curve. The solution of the equation $P_0 D_r^*(\bar{Q}, \bar{g}, g_{R,0}, \psi_r) = 0$ in ψ_r equals

$$
\begin{aligned}
\psi_r &= \frac{P_0 \left\{ \frac{I(R=r)-I(R=0)}{g_0(R|W)} (Y - \bar{Q}(W, R)) + \bar{Q}(W, r) - \bar{Q}(W, 0) \right\}}{P_0 \left\{ \frac{I(R=r)-I(R=0)}{g_0(R|W)} (A - \bar{g}(W, R)) + \bar{g}(W, r) - \bar{g}(W, 0) \right\}} \\
&= \frac{P_0 \left\{ \bar{Q}_0(W, r) - \bar{Q}_0(W, 0) \right\}}{P_0 \left\{ \bar{g}_0(W, r) - \bar{g}_0(W, 0) \right\}} \\
&= \psi_{r,0}.
\end{aligned}
$$

Thus this solution is correct even if both $\bar{Q}$ and $\bar{g}$ are misspecified. This result also implies a robustness for $D^* = \sum_{r>0} h(r) D_r^*$.

TMLE of Ψ. Define

$$
\begin{aligned}
C_{Y,r}(P) &= \frac{1}{E\{A(r) - A(0)\}} \frac{I(R = r) - I(R = 0)}{g(R \mid W)} \\
C_{A,r}(P) &= \frac{E(Y(r) - Y(0))}{E^2(A(r) - A(0))} \frac{I(R = r) - I(R = 0)}{g(R \mid W)}.
\end{aligned}
$$

If Y is continuous in $(0, 1)$ or binary in $\{0, 1\}$, then we can use the quasi-log-likelihood loss function $L_Y(\bar{Q})(O) = Y \log \bar{Q}(W, R) + (1 - Y) \log(1 - \bar{Q}(W, R))$ for $\bar{Q}_0$. Regarding parametric submodel $\bar{Q}(\epsilon)$, we use the logistic regression $\text{logit}\bar{Q}(\epsilon) = \text{logit}\bar{Q} + \epsilon C_Y$ with clever covariate $C_Y = \sum_{r>0} h(r) C_{Y,r}$.

Similarly, if A is continuous in $(0, 1)$ or binary in $\{0, 1\}$, then we can use the quasi-log-likelihood loss function $L_A(\bar{g})(O) = A \log \bar{g}(W, R) + (1 - A) \log(1 - \bar{g}(W, R))$ for $\bar{g}_0$. Regarding parametric submodel $\bar{g}(\epsilon)$, we use the logistic regression $\text{logit}\bar{g}(\epsilon) = \text{logit}\bar{g} + \epsilon C_A$ with clever covariate $C_A = \sum_{r>0} h(r) C_{A,r}$. For the marginal distribution of W, we use the log-likelihood loss function $L_W(Q_W) = -\log Q_W$, and as submodel we select $(1 + \epsilon D_W^*) Q_W$, where $D_W^* = \Pi(D^* \mid T_W)$. We can now define the loss function $L(Q_W, \bar{Q}, \bar{g}) = L_W(Q_W) + L_Y(\bar{Q}) + L_A(\bar{g})$ for the combined $(Q, \bar{g})$, and the submodel above was selected so that $\frac{d}{d\epsilon} L(Q_W(\epsilon_1), \bar{Q}(\epsilon_2), \bar{g}(\epsilon_2))$ at $\epsilon = 0$ spans the efficient influence curve $D^*(Q, \bar{g})$. Here $Q = (Q_W, \bar{Q})$.

The initial estimator of Q_W is the empirical distribution function. We can obtain initial estimators of $\bar{Q}_0$, $\bar{g}_0$, and $g_{R,0}$ with loss-based learning. Let $(Q_n^0, \bar{g}_n^0, g_{n,R})$ represent this initial estimator of $(Q_{W,0}, \bar{Q}_0, \bar{g}_0, g_{0,R})$. The TMLE is now defined: $\epsilon_n^1 = \arg\min_\epsilon P_n L(Q_n^0(\epsilon), \bar{g}_n^0(\epsilon))$, set $Q_n^1 = Q_n^0(\epsilon_n^1)$, $\bar{g}_n^1 = \bar{g}_n^0(\epsilon_n^1)$, and iterate this updating process to convergence. We note that ϵ_1 is estimated at zero so that the empirical distribution of W will not be updated, and $g_{n,R}$ is not updated either. Let $Q_n^*, \bar{g}_n^*$ denote the limit. Then the TMLE of ψ_0 is given by $\psi_{1n}^* = \Psi_1(Q_n^*, \bar{g}_n^*)$.

In this formulation of the TMLE, we are not providing a guarantee that ψ_n^* is a completely valid substitution estimator since the conditional means $\bar{Q}_n^*(W, R)$ and $\bar{g}_n^*(W, R)$ are not variation independent. The formal recipe of TMLE can be based on the orthogonal factorization of the density $P(O) = P(W)P(R \mid W)P(A \mid W, R)P(Y \mid W, R, A)$ in variation-independent conditional distributions, providing a loss function for $\bar{Q}(W, R, A)$ (instead of $\bar{Q}(W, R)$), the required parts of the conditional distribution of A, given W, R, and for the marginal distribution of W, and choosing working submodels based on the corresponding orthogonal decomposition of the efficient influence curve. We leave this exercise to the reader.

A.15 Example: TMLE of the Conditional Relative Risk

We consider n i.i.d. observations of $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where W is a vector of baseline covariates, A is an exposure of interest, and $Y = \{0, 1\}$ is a binary outcome. We define the statistical model $\mathcal{M}$ as all probability distributions P_0 satisfying

$$\bar{Q}_0(A, W) = e^{m_{\beta_0}(A,V)}\theta_0(W),$$

where $\bar{Q}_0(A, W) \equiv P_0(Y = 1 \mid A, W)$, $m_{\beta_0}(A, V)$ is a specified function of A and effect modifiers $V \subset W$, and $\theta_0(W) \equiv P_0(Y = 1 \mid A = 0, W)$. We will also use the notation $\bar{Q}_{\beta_0,\eta_0}$ for $\bar{Q}_0$. For simplicity, we first consider the case where $m_{\beta_0}(A, V) = \beta_0 A$, but we also provide the general formulas below.

Constructing the efficient score. The probability distribution of O in this semiparametric model is indexed by a finite-dimensional parameter β and infinite-dimensional nuisance parameter η consisting of θ, the marginal distribution of W, and the conditional distribution of A, given W. Let g_0 denote the conditional distribution of A, given W. The efficient influence curve $D^*(P_0)$ at P_0 happens to only depend on P_0 through β_0, θ_0, and g_0, so that we will also denote it by $D^*(\beta_0, \eta_0)$ or $D^*(\beta_0, \theta_0, g_0)$. We have

$$D^*(\beta_0, \eta_0) = -\left[\tfrac{d}{d\beta_0} P_0 S^*(\beta_0, \eta_0)\right]^{-1} S^*(\beta_0, \eta_0), \tag{A.8}$$

where $S^*(\beta_0, \eta_0)$ denotes the efficient score given by $S(\beta_0, \eta_0) - \Pi(S(\beta_0, \eta_0) \mid T_{nuis})$. Here $S(\beta_0, \eta_0)(Y \mid A, W) = \frac{d}{d\beta_0} \log P_{\beta_0,\theta_0}(Y \mid A, W)$ is the score of the parameter of interest β_0, and T_{nuis} is the nuisance tangent space, viewed as a subspace of the Hilbert space $L_0^2(P_0)$ endowed with the inner product $\langle h_1, h_2 \rangle = E_0 h_1 h_2(O)$. Recall

that a projection of a function S on a subspace T_{nuis} of a Hilbert space is uniquely defined as follows: (1) the projection is an element of the subspace T_{nuis}, and (2) $S - \Pi(S \mid T_{nuis}) \perp T_{nuis}$. T_{nuis} is the direct sum of the three orthogonal spaces involving each of the nuisance parameters: $T_{nuis} = T_W \bigoplus T_{A|W} \bigoplus T_\theta$. Specifically, T_W consists of all functions in $L_0^2(P_0)$ of W with mean zero, $T_{A|W}$ consists of all functions in $L_0^2(P_0)$ of (A, W) with conditional mean zero, given W, and T_θ is the tangent space spanned by all the scores of parametric submodels through P_0 that only fluctuate θ. Thus,

$$S^*(\beta_0, \eta_0) = S(\beta_0, \eta_0) - [\Pi(S(\beta_0, \eta_0) \mid T_W) + \Pi(S(\beta_0, \eta_0) \mid T_{A|W}) + \Pi(S(\beta_0, \eta_0) \mid T_\theta)].$$

We have $\log P_{\beta,\theta}(Y = 1 \mid A, W) = \log \theta(W) + \beta A$. It follows that

$$S(\beta_0, \eta_0)(O) = \frac{d}{d\beta_0} \log P_{\beta_0,\theta_0}(Y \mid A, W) = \frac{A}{1-\bar{Q}_{\beta_0,\theta_0}}(Y - \bar{Q}_{\beta_0,\theta_0}(A, W)).$$

Since $S(\beta_0, \eta_0)$ has a conditional mean, given (A, W), equal to zero, it follows that it is orthogonal to T_W and $T_{A|W}$, so that its projection onto $T_W + T_{A|W}$ equals zero.

To calculate the tangent space T_θ, we consider submodels $P_0(\epsilon)(Y \mid A, W)$ implied by $\log \bar{Q}_0(\epsilon)(A, W) = \log \theta_0(W) + \beta_0 A + \epsilon h_3(W)$ for an arbitrary function h_3. Notice that this indeed implies a submodel in our semiparametric regression model. It is straightforward to show that the score of this submodel at $\epsilon = 0$ equals $1/(1 - \bar{Q}_0(A, W))h_3(W)(Y - \bar{Q}_0(A, W))$. This shows that $T_\theta = \{1/(1 - \bar{Q}_0(A, W))h_3(W)(Y - \bar{Q}_0(A, W)) : h_3\}$.

It remains to determine $\Pi(S(\beta_0, \eta_0) \mid T_\theta)$. As repeatedly used and shown in van der Laan and Robins (2003), any function $S(B, Pa(B))$ of a binary variable B and other variables $Pa(B)$ that has a conditional mean of zero, given $Pa(B)$, can be written as $(S(1, Pa(B)) - S(0, Pa(B))(B - P(B = 1 \mid Pa(B)))$. For a function V, let $h_V(A, W) = (V(1, A, W) - V(0, A, W))$, so that $V - E_0(V \mid A, W) = h_V(A, W)(Y - \bar{Q}_0)$. Thus,

$$\begin{aligned}\Pi(V \mid T_\theta) &= \Pi(V - E_0(V \mid A, W) \mid T_\theta) \\ &= \Pi(h_V(Y - \bar{Q}_0) \mid T_\theta).\end{aligned}$$

We need to find h_3^* such that

$$\begin{aligned}E_0\left[\left\{h_V(A, W)(Y - \bar{Q}_0) - \frac{h_3^*(W)}{1 - \bar{Q}_0}(Y - \bar{Q}_0)\right\}\frac{h_3(W)}{1 - \bar{Q}_0}(Y - \bar{Q}_0)\right] &= 0 \text{ for all } h_3(W), \\ E_0\left[\left(h_V(A, W) - \frac{h_3^*(W)}{1 - \bar{Q}_0}\right)\frac{h_3(W)}{1 - \bar{Q}_0}(Y - \bar{Q}_0)^2\right] &= 0 \text{ for all } h_3(W), \\ E_0\left[\left(h_V(A, W) - \frac{h_3^*(W)}{1 - \bar{Q}_0}\right)\frac{h_3(W)}{1 - \bar{Q}_0}\sigma^2(A, W)\right] &= 0 \text{ for all } h_3(W), \\ E_0\left[\left(h_V(A, W)\frac{\sigma^2}{1 - \bar{Q}_0} - \frac{h_3^*(W)}{(1 - \bar{Q}_0)^2}\sigma^2\right)h_3(W)\right] &= 0 \text{ for all } h_3(W),\end{aligned}$$

$$E_0\left[\left(E_0\left[\frac{h_V(A,W)}{1-\bar{Q}_0}\sigma^2 \mid W\right] - h_3^*(W)E_0\left[\frac{\sigma^2}{(1-\bar{Q}_0)^2} \mid W\right]\right)h_3(W)\right] = 0 \text{ for all } h_3(W),$$

where $\sigma^2(A,W) = \mathrm{VAR}_0(Y \mid A,W) = \bar{Q}_0(1-\bar{Q}_0)$. Therefore

$$h_3^*(W) = \frac{E_0\left(\frac{h_V(A,W)\sigma^2}{1-\bar{Q}_0} \mid W\right)}{E_0\left(\frac{\sigma^2}{(1-\bar{Q}_0)^2} \mid W\right)}.$$

This provides us with the projection of V onto the nuisance tangent space T_{nuis}. In particular, if $V = S(\beta_0,\eta_0)$, we have $V = A/(1-\bar{Q}_0)(Y-\bar{Q}_0)$, so that $h_V = A/(1-\bar{Q}_0)$. This yields

$$\Pi(S(\beta_0,\eta_0) \mid T_\theta) = \frac{E_0\left[\frac{A\bar{Q}_0}{1-\bar{Q}_0} \mid W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0} \mid W\right]}\frac{(Y-\bar{Q}_0)}{1-\bar{Q}_0}.$$

It follows that the efficient score is given by

$$S^*(\beta_0,\eta_0) = \left(A - \frac{E_0\left[\frac{A\bar{Q}_0}{(1-\bar{Q}_0)} \mid W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0} \mid W\right]}\right)\frac{(Y-\bar{Q}_0)}{1-\bar{Q}_0}. \tag{A.9}$$

Double robustness of efficient score. It is of interest to note that the efficient score can also be represented as:

$$S^*(\beta_0,\eta_0) = h^*(A \mid W)\left(Y\frac{\bar{Q}_0(0,W)}{\bar{Q}_0(A,W)} - \bar{Q}_0(0,W)\right),$$

where

$$h^*(A \mid W) \equiv \frac{\bar{Q}_0}{\bar{Q}_0(0,W)(1-\bar{Q}_0)}\left(A - \frac{E_0\left[\frac{A\bar{Q}_0}{(1-\bar{Q}_0)}\|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}\right)$$

is a function satisfying $E_0(h^*(A \mid W) \mid W) = 0$. This representation shows that $P_0S^*(\beta_0,\theta,g) = 0$ if either $\theta = \theta_0$ or $g = g_0$, thereby establishing the double robustness of the efficient score as the estimating function for β_0.

The derivation above assumes that $m_{\beta_0}(A,V) = \beta_0 A$. In general, the efficient score is given by

$$S^*(\beta_0,\eta_0)(O) = \frac{1}{1-\bar{Q}_0}\left(\frac{d}{d\beta_0}m_{\beta_0} - \frac{E_0\left[\frac{d}{d\beta_0}m_{\beta_0}\frac{\bar{Q}_0}{(1-\bar{Q}_0)} \mid W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0} \mid W\right]}\right)(Y-\bar{Q}_0).$$

The efficient influence curve is defined as the standardized version $c_0^{-1}S^*(\beta_0,\eta_0)$, where $c_0 = -\frac{d}{d\beta_0}P_0S^*(\beta_0,\eta_0)$.

Constructing a parametric submodel having a score that spans the efficient score. If we assume $\bar{Q}_0 = \exp(m_{\beta_0}(A,W))\theta_0(W)$, and we use as submodel $\log\bar{Q}_0(\epsilon) = m_{\beta_0+\epsilon} + \log\theta_0 + \epsilon r$, then the score equals $(d/d\beta_0 m_{\beta_0} + r)(Y-\bar{Q}_0)/(1-\bar{Q}_0)$. Thus, to

arrange that this score equals the efficient score we set r equal to

$$r^*(\bar{Q}_0, g_0) = -\frac{E_0\left[d/d\beta_0 m_{\beta_0} \frac{\bar{Q}_0}{(1-\bar{Q}_0)} \mid W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0} \mid W\right]}.$$

The iterative TMLE. This defines the desired ϵ-extension $\bar{Q}_n^0(\epsilon)$ of an initial fit $\bar{Q}_n^0$. We use the log-likelihood loss function for $\bar{Q}_0$: $L(\bar{Q}) = -\{Y \log \bar{Q} + (1 - Y) \log(1 - \bar{Q})\}$. For example, if $m_\beta(W, A) = \beta A$, then this ϵ-fluctuation corresponds with adding $\epsilon C(A, W)$ to the initial fit $\log \bar{Q}_n^0(A, W) = \beta_n^0 A + r_n^0(W)$, where the clever covariate is given by

$$C(Q_n^0, g_n^0)(A, W) = A - \frac{E_{g_n^0}\left(\frac{Q_n^0(A,W)}{1-Q_n^0(A,W)} A \mid W\right)}{E_{g_n^0}\left(\frac{Q_n^0(A,W)}{(1-Q_n^0)(A,W)} \mid W\right)}.$$

Let ϵ_n^0 be the maximum likelihood estimator over ϵ for this parametric submodel $\{Q_n^0(\epsilon) : \epsilon\}$. This requires fitting a log-binomial regression model. Let $\bar{Q}_n^1 = \bar{Q}_n^0(\epsilon_n^0)$ be the updated estimate of $\bar{Q}_0$, which corresponds with an updated β_n^1 and θ_n^1. We iterate this updating process until the corresponding sequence β_n^k is such that $\beta_n^k - \beta_n^{k-1}$ no longer significantly change. We denote the selected final update by $\bar{Q}_n^*$ and let β_n^* be the corresponding TMLE of β_0.

A.16 IPCW Reduced-Data TMLE

Summary. IPCW estimators have gained popularity due to their simplicity. However, this gain in simplicity comes at a severe cost in terms of bias and variance. We show that by inverse probability of censoring weighting a TMLE based on a reduction of the original observed data structure, one obtains a valid substitution estimator of the target parameter. This estimator is a special case of the TMLE presented in Appendix A.6, corresponding with a particular IPCW loss function and parametric fluctuation function. These estimators are relatively easy-to-implement substitution estimators with good efficiency and robustness properties.

The TMLE of a target parameter ψ_0 is characterized by two ingredients: a choice of loss function for Q_0 and a parametric fluctuation working model to fluctuate Q. These two choices combined determine the estimating function $D(Q, \eta) \equiv \frac{d}{d\epsilon} L(Q_\eta(\epsilon))\Big|_{\epsilon=0}$ whose estimating equation $P_n D(Q_n^*, \eta_n) = 0$ will be solved by the resulting TMLE Q_n^*. If D is the efficient influence curve, then we will refer to this TMLE as an efficient TMLE. The efficient TMLE, based on, e.g., the log-likelihood loss function $L(Q)$ and efficient influence curve estimating function $D^*()$, can be quite involved for complex longitudinal data structures with time-dependent covariates, since Q_0 may be a very high-dimensional function. Therefore, it is of interest

to also provide TMLE for which Q_0 is chosen to be of lower dimension, at the cost of having to work with a loss function $L_{\eta_0}(Q)$ that is indexed by an unknown nuisance parameter, and fluctuation model that generates an inefficient estimating function $D()$. For that purpose we propose a general class of so-called inverse probability of censoring-weighted reduced-data TMLEs, which modify the efficient TMLE for a user-supplied reduced (simplified-)data structure by weighting the loss function with inverse probability of censoring weights.

Let $O = (L(0), A(0), \ldots, L(K), A(K), L(K+1)) \sim P_0$. Assume an SCM $A(t) = f_{A(t)}(Pa(A(t)), U_{A(t)})$, $t = 1, \ldots, K$, $L(t) = f_{L(t)}(Pa(L(t)), U_{L(t)})$, $t = 1 \ldots, K+1$, where $Pa(A(t)) = (\bar{A}(t-1), \bar{L}(t))$, and $Pa(L(t)) = (\bar{A}(t-1), \bar{L}(t-1))$. Here $A(t)$, $t = 0, \ldots, K$ denote the intervention nodes, which can include both treatment and censoring actions. This SCM allows us to define counterfactuals L_a and L_d indexed by static interventions a and dynamic treatments d, respectively. We assume the SRA about the error nodes U in the SCM so that the g-computation formula provides us with the identifiability of any parameter of the distribution of a counterfactual L_d for a given rule d, possibly a static rule. Specifically, under this SRA, the probability distribution of the observed data random variable $O = (A, L = L_A)$ factorizes into a factor Q_0 implied by the full-data distribution of the counterfactuals $X = (L_a : a)$ and a factor $g_0(\cdot \mid X) = \prod_{t=0}^{K} g_{A(t),0}(A(t) \mid Pa(A(t)))$ that corresponds with the conditional distribution of A, given X:

$$P_{Q_0,g_0}(O) = \prod_{t=0}^{K+1} Q_{L(t),0}(L(t) \mid Pa(L(t)) \prod_{t=0}^{K} g_{A(t),0}(A(t) \mid Pa(A(t))).$$

By SRA (which implies coarsening at random), we have $Q_{L(t),0}(l(t) \mid \bar{l}(t-1), \bar{a}(t-1)) = P(L_a(t) = l(t) \mid \bar{L}_a(t-1) = \bar{l}(t-1))$ so that indeed Q_0 represents the identifiable part of the full-data distribution of the counterfactuals X.

A statistical model $\mathcal{M}$ for P_0 can be represented as all probability distributions $P_{Q,g}$ with $Q \in \mathcal{Q}$ and $g \in \mathcal{G}$ for some specified models $\mathcal{Q}$ and SRA model $\mathcal{G}$ for Q_0 and g_0, respectively. Given a parameter $\Psi : \mathcal{Q} \to \mathbb{R}^d$, our goal is to estimate $\Psi(Q_0)$.

The basic idea of IPCW-R-TMLE is as follows. Our target parameter can also be written as a function of the distribution of a reduction L_a^r of the counterfactual L_a, obtained by removing a number of the time-dependent components of $L_a(t)$. Thus, we can write our parameter as $\Psi(Q_0) = \Psi^r(Q_0^r)$, where $Q_0^r(a, l^r)$ represents the distribution of L_a^r: $Q^r_{L^r(t),0}(l^r(t) \mid \bar{l}^r(t-1), \bar{a}(t-1)) = P(L_a^r(t) = l^r(t) \mid \bar{L}_a^r(t-1) = \bar{l}^r(t-1))$. Now, we note that the inverse-weighted log-likelihood loss function, for any marginal probability distribution g^r of A, $L_{g_0}(Q^r) \equiv -g^r/g_0 \log Q^r$, is a valid loss function for Q_0^r, since

$$Q^r \to P_0 \log Q^r \frac{g^r}{g_0} = P_{Q_0,g^r} \log Q^r = P_{Q_0^r,g^r} \log Q^r,$$

is maximized at Q_0^r. To see the last equality, use the representation $O = (A, L_A)$ and that $Q_0(a, l) = P_0(L_a = l)$ so that

$$E_{Q_0,g^r} \log Q^r(A, L_A^r) = E_{Q_0^r} \sum_a \log Q^r(a, L_a^r) g^r(a),$$

which is indeed maximized at Q_0^r. In addition, the inverse-probability-weighted reduced-data efficient influence curve, $D(Q^r, g^r, g_0) \equiv D^{*r}(Q^r, g^r)g^r/g_0$, is a targeted estimating function for the target parameter $\Psi(Q_0^r)$, as discussed in detail below. We can now apply TMLE, as described in the Appendix A.6, with this inverse-weighted log-likelihood loss function, a fluctuation working model $\{Q_{g^r}^r(\epsilon) : \epsilon\}$ with score at $\epsilon = 0$ equal to the reduced-data efficient influence curve $D^{r*}(Q^r, g^r)$, so that the TMLE will solve $P_n D(Q_n^{r*}, g^r, g_n) = 0$. Our proposal below refines the choice of IPCW log-likelihood loss function by inverse weighting each factor $Q_{L^r(t),0}$ of Q_0^r separately with more stable weights $g^r(\bar{A}(t-1) \mid X^r)/g_0(\bar{A}(t-1) \mid X)$, as described in the procedure below.

We will use the notation $\mathcal{M}(g) = \{P_{Q,g} : Q \in \mathcal{Q}\}$ for the statistical model implied by a model $\mathcal{Q}$ for Q_0 and a treatment mechanism g contained in the set $\mathcal{G}$ of all SRA-conditional distributions of A, given X. We note that, since Q_0 is identifiable based on i.i.d. sampling from an element in $\mathcal{M}(g)$, one can also view Ψ as a parameter on the model $\mathcal{M}(g)$. The IPCW-R-TMLE is defined by the following steps.

(Optional) specify reduced-data structure. Determine a reduction $O^r = (A, L^r)$ of $O = (A, L)$, where L^r is a function of L and where the reduction is such that it is still possible to identify the parameter of interest ψ_0 from the probability distribution of $O^r = (A, L^r = L_A^r)$ under the SRA for the reduced full-data structure $X^r = (L_a^r : a \in \mathcal{A})$. In other words, $\Psi(Q_0)$ needs to depend on the distribution of $X = (L_a : a)$ only through the distribution of $X^r = (L_a^r : a)$. For example, $O = (W = L(0), A, \bar{L}(K), Y = L(K+1))$ consists of baseline covariates W, treatment regimen $A = (A(0), \ldots, A(K))$, time-dependent covariate process $\bar{L}(K)$, and a final outcome Y, one is concerned with the estimation of EY_a for some static regimen a, and one defines $O^r = (W, A, Y)$, which is obtained from O by deleting all time-dependent covariates.

Reduced-data model. Let $O^r = (A, L_A^r)$, $X^r = (L_a^r : a)$, g^r a conditional distribution of A, given X^r, satisfying SRA with respect to reduced-data O^r. Let $\mathcal{M}^r(g^r) = \{P_{Q^r,g^r}^r = Q^r g^r : Q^r \in \mathcal{Q}^r\}$ be a statistical model for O^r, where the model $\mathcal{Q}^r = \{Q^r : Q \in \mathcal{Q}\}$ for Q_0^r is implied by the model $\mathcal{Q}$ for Q_0. Let $\Psi^r : \mathcal{Q}^r \to \mathbb{R}^d$ be such that $\Psi^r(Q^r) = \Psi(Q)$ for all $Q^r \in \mathcal{Q}^r$, and, in particular, $\Psi^r(Q_0^r) = \Psi(Q_0)$. In the example with $O^r = (W, A, Y)$, g^r is a conditional distribution of A, given W, Q^r is the distribution of $(W, (Y_a : a))$ implied by the distribution of L_a under Q, and $\Psi^r(Q^r) = E_{Q^r} Y_a$. In particular, if the data are not reduced in the previous step, then $O^r = O$, $Q^r = Q$, $g^r = g$, $\mathcal{M}^r(g^r) = \mathcal{M}(g)$, $\Psi^r = \Psi$.

Factorization of Q^r. Suppose $P_{Q_0^r,g_0^r} = \prod_j Q_{0j}^r g_0^r$ factors into various terms Q_{0j}^r, $j = 1, \ldots, J$ (e.g., $J = K+1$). Suppose that $Q_{0j}^r(O^r)$ depends on O^r only through $(A(0), \ldots, A(j^r - 1), \bar{L}^r(j^r))$, $j = 1, \ldots, J$. In a typical scenario, we have that Q_{0j}^r denotes the conditional distribution of $L^r(j^r)$, given $(A(0), \ldots, A(j^r - 1))$ and $\bar{L}^r(j^r - 1)$. For notational convenience, we used the short-hand notation $j^r = j^r(j)$, suppressing its deterministic dependence on j. In the example with $O^r = (W, A, Y)$, Q_0^r factors as $Q_0^r(w, a, y) = Q_{0w}(w) Q_0^r(y \mid w, a)$, where $Q_0^r(y \mid w, a) = P_0(Y(a) = y \mid W = w)$, giving us factorization $Q^r(w, a, y) = Q_1^r(w) Q_2^r(y \mid w, a)$. In

particular, if the data are not reduced, then $P_{Q_0,g_0} = \prod_t Q_{t0}g_0$, $t = 1, \ldots, K+1$, where $Q_{t,0}$ denotes the conditional distribution of $L(t)$, given $\bar{L}(t-1), \bar{A}(t-1)$, so that $Q_{t,0}(O)$ depends on O only through $(A(0), \ldots, A(t-1))$, $t = 1, \ldots, K+1$.

Determine Q_j^r-components of efficient influence curve for reduced-data model. Let $D^r(P^r)$ be the efficient influence curve at $P^r = P^r_{Q^r,g^r} = Q^r g^r$ for the parameter Ψ^r in the model $\mathcal{M}^r(g^r)$ for the reduced-data structure O^r. This efficient influence curve can be decomposed orthogonally as $D^r(P^r) = D^r(Q^r, g^r) = \sum_{j=1}^J D_j^r(P^r)$, where $D_j^r(P^r)$ is an element of the tangent space generated by the jth factor Q_j^r of $Q^r = \prod_j Q_j^r$ at P^r, $j = 1, \ldots, J$. In the example with $O^r = (W, A, Y)$, this efficient influence curve for the reduced data is given by (and decomposed as):

$$D^r(Q^r, g^r)(W, A, Y) = \{I(A = a)/g^r(a \mid W)(Y - \bar{Q}^r(A, w))\} + \{\bar{Q}^r(a, W) - \Psi^r(Q^r)\},$$

where $\bar{Q}^r(a, w) = E_{Q^r}(Y_a \mid W)$. This defines D_1^r and D_2^r. In particular, if the data were not reduced and the model for Q_0 is nonparametric, then the efficient influence curve $D(P) = \sum_{t=1}^{K+1} D_t(P)$, with

$$D_t(P) = E_P(D(P)(O) \mid \bar{A}(t-1), \bar{L}(t)) - E_P(D(P)(O) \mid \bar{A}(t-1), \bar{L}(t-1))$$

being the projection of $D(P)$ on the tangent space generated by the conditional distribution Q_t of $L(t)$, given $\bar{L}(t-1), \bar{A}(t-1)$.

Determine hardest Q_j^r-fluctuation working models. Given a Q^r, construct submodels $\{Q_j^r(\epsilon) : \epsilon\}$ through Q_j^r and loss functions $L(Q_j^r)$, such as $L(Q_j^r) = -\log Q_j^r$, so that

$$\frac{d}{d\epsilon} L(Q_j^r(\epsilon))\Big|_{\epsilon=0} = D_j^r(Q^r, g^r), \ j = 1, \ldots, J.$$

In the example, [say Y is binary or bounded in $(0, 1)$] we can fluctuate Q_2^r using a logistic fluctuation working model with clever covariate $I(A = a)/g^r(A \mid W)$ and employ the binary-outcome log-likelihood loss function for the conditional mean (or probability) $\bar{Q}_2^r(w, a) = E_{Q_2^r}(Y_a \mid W = w)$. In particular, if the data are not reduced, then, given a $Q \in \mathcal{Q}$ construct submodels $\{Q_t(\epsilon) : \epsilon\}$ through Q_t at $\epsilon = 0$, with score at $\epsilon = 0$ equal to $D_t(Q, g)$,

$$\frac{d}{d\epsilon} L(Q_t(\epsilon))\Big|_{\epsilon=0} = D_t(Q, g), \ t = 1, \ldots, K+1.$$

Construct IPCW weights for each j-specific Q_j^r-factor. For each j, construct the weight function

$$w_{j,0} = \frac{g_0^r(\bar{A}(j^r - 1) \mid X^r)}{g_0(\bar{A}(j^r - 1) \mid X)}, \ j = 1, \ldots, J.$$

We will often denote the weights $g_0^r(\bar{A}(j^r - 1) \mid X^r)/g_0(\bar{A}(j^r - 1) \mid X)$ by $g_{j,0}^r/g_{j,0}$. We note that for each $j = 1, \ldots, J$

$$Q_{j,0}^r = \arg \min_{Q_j^r \in \mathcal{Q}_j^r} P_{Q_0,g_0} w_{j,0} L(Q_j^r) = \arg \min_{Q_j^r \in \mathcal{Q}_j^r} P_{Q_0^r,g_0^r} L(Q_j^r),$$

such that, for each choice g_0^r (can be any function), the IPCW-R loss function $L_{w_0}(Q^r) \equiv \sum_j w_{j,0} L(Q_j^r)$ is a valid loss function for Q_0^r [i.e., for the true distribution of $(L_a^r : a)$], indexed by nuisance parameter (g_0^r, g_0). In our example we have $w_{1,0} = 1$ (no weighting for marginal distribution of W) and $w_{2,0} = g_0^r(A \mid W)/g_0(A \mid X)$, and the IPTW-R loss function for Q^r is $\sum_{j=1}^2 L(Q_j^r) w_{j,0}$. Note that $g_0^r(A \mid W) = \prod_{t=0}^K g_0^r(A(t) \mid \bar{A}(t-1), W)$. If the data are not reduced, then $w_{t,0} = 1$, and one could select $L(Q) = -\log Q$ as the log-likelihood loss function.

Estimate the weights. Construct estimators g_n^r and g_n of g_0^r and g_0, respectively, and construct the corresponding estimator w_n of the weight function w_0. In our example, this requires fitting the conditional distribution of the time-dependent treatments $A(t)$, given past treatment, and baseline covariates (thus ignoring the time-dependent covariates), as well as the true treatment mechanism.

IPCW-R-TMLE at specified weights. We will now compute the TMLE under i.i.d. sampling $O_1^r, \ldots, O_n^r$ from $P^r_{Q_0^r, g^r}$, but assigning the above IPCW weights w_n, as follows. Let $Q^{r,0}$ be an initial estimator of Q_0^r. For example, let $Q_j^{r,0} = \arg\min_{Q_j^r \in \mathcal{Q}_j^r} \sum_i L(Q_j^r)(O_i^r) w_{j,i,n}$ be a weighted maximum likelihood estimator of Q_{0j}^r according to a working model $\mathcal{Q}_j^r$. In general, we can use a weighted-ML-based estimator, such as the super learner, based on this weighted log-likelihood loss function $L_{w_0}(Q^r) = \sum_{j=1}^J L(Q_j^r) w_{j,0}$. Subsequently, we compute the overall amount of fluctuation with an IPCW loss-based estimation,

$$\epsilon_n^1 = \arg\min_\epsilon \sum_i \sum_j L(Q_j^{r,0}(\epsilon))(O_i^r) w_{j,i,n},$$

and compute the corresponding first-step targeted update $Q_j^{r,1} = Q_j^{r,0}(\epsilon_n^1)$, $j = 1, \ldots, J$, and thereby the overall update $Q^{r,1} = Q^{r,0}(\epsilon_n^1)$. Iterate this process till convergence (i.e., $\epsilon_n^k \approx 0$) and denote the final update by $Q_n^{r*} = (Q_{j,n}^{r*} : j = 1, \ldots, J)$. Let $D(Q^r, g^r, g) = \sum_j D_j^r(Q^r, g^r) \frac{g_j^r}{g_j}$ be the IPCW efficient influence curve estimating function for the reduced-data structure O^r. We have that Q_n^{r*}, in conjunction with an estimate of the weights, solves the corresponding estimating equation:

$$0 = \sum_i D(Q_n^{r*}, g_n^r, g_n)(O_i) = \sum_i \sum_j D_j^r(Q_n^{r*}, g_n^r)(O_i^r) w_{j,i,n}. \tag{A.10}$$

In our example, the marginal empirical distribution of W would not be updated (we would use separate ϵ for fluctuation of the marginal distribution of W), so that only $Q_2^{r,0}$ is updated, and the IPCW-R efficient influence curve is given by $D_1^r(W) + D_2^r(W, A, Y) g_0^r(A \mid W)/g_0(A \mid X)$. In particular, if the data are not reduced, then Q_n^*, g_n solves the efficient influence curve equation $0 = \sum_i \sum_t D_t(Q_n^*, g_n)(O_i)$.

Substitution estimator. Our estimator of ψ_0 is given by $\Psi^r(Q_n^{r*})$. In our example, EY_a is estimated as $\Psi^r(Q_n^{r*}) = \sum_w Q_{1,n}^r(w) \sum_y y Q_{2,n}^{r*}(y \mid w, a)$. In particular, if the data are not reduced, then ψ_0 is estimated with $\Psi(Q_n^*)$.

The IPCW-R-TMLE is an estimator Q_n^{r*} of Q_0^r (i.e., of the true distribution of L_a^r for each a), solving an IPCW reduced-data efficient influence curve equation (A.10).

Firstly, we establish that this IPCW reduced-data efficient influence curve is an "estimating function" for the target parameter $\Psi^r(Q_0^r)$ with nice robustness properties with respect to its nuisance parameters Q_0^r and g_0 (for each choice of g_0^r). Subsequently, we discuss the corresponding implications for the statistical properties of the IPCW-R-TMLE.

Robustness properties of IPCW reduced-data efficient influence function. Recall that $D^r(Q^r, g^r)$ denotes the efficient influence curve for the reduced-data $O^r \sim P_{Q^r,g^r}$ for model $\mathcal{M}^r$ and parameter Ψ^r. It follows from the general results in van der Laan and Robins (2003) that $P_{Q_0^r,g_0^r} D^r(Q^r, g^r) = 0$ if either $Q^r = Q_0^r$ or $\Psi(Q^r) = \Psi(Q_0^r)$ and $g^r = g_0^r$. This double robustness result for D^r is exploited/inherited by the estimating function $D(Q^r, g^r, g_0) \equiv \sum_j D_j^r(Q^r, g^r) g_j^r / g_{0j}$, whose corresponding estimating equation is solved by our IPCW-R-TMLE, in the following manner. If the denominator of the weights $g = g_0$ is correctly specified, then we have

$$P_{Q_0,g_0} D(Q^r, g^r, g_0) = P_{Q_0,g_0} \sum_j D_j^r(Q^r, g^r) \frac{g_j^r}{g_{j,0}} = P_{Q_0,g^r} \sum_j D_j^r(Q^r, g^r).$$

This implies that if $g = g_0$ (i.e., the action mechanism is correctly specified), then $P_{Q_0,g_0} D(Q^r, g^r, g_0) = 0$ for all choices of Q^r, g^r with $\Psi(Q^r) = \Psi(Q_0^r)$. That is, $D(Q^r, g^r, g_0)$ represents an unbiased estimating function in ψ for each choice of g^r.

In a typical scenario, we have that $Q_{j,0}^r$ denotes the conditional distribution of $L^r(j^r)$, given $A(0), \ldots, A(j^r - 1)$ and $\bar{L}^r(j^r - 1)$. In this case, if $g_{j,0}$ is only a function of O^r, then if $Q^r = Q_0^r$, it follows that $P_{Q_0,g^r} D_j^r(Q_0^r, g^r) \frac{g_j^r}{g_j} = 0$ for all g_j only being a function of O^r [by using that the conditional expectation of a score $D_j^r(Q_0^r, g^r)$ of Q_{j0}^r, given $(A(0), \ldots, A(j^r - 1)$ and $\bar{L}^r(j^r - 1)$, equals zero], and as a consequence, $P_{Q_0,g_0} D(Q_0^r, g^r, g) = 0$ for such misspecified g. That is, in the case that the true g_0 and its asymptotic (possibly misspecified) fit are only functions of the reduced-data structure O^r, we have the double robustness of the estimating function $D(Q^r, g^r, g)$ in the sense that, for any choice g^r, $P_{Q_0,g_0} D(Q^r, g^r, g) = 0$ if $\Psi(Q^r) = \Psi(Q_0^r)$ and, either $Q^r = Q_0^r$ or $g = g_0$. In particular, if the data are not reduced, then we have $P_{Q_0,g_0} D(Q, g) = 0$ if $\Psi(Q) = \psi_0$ and either $Q = Q_0$ or $g = g_0$. In fact, the efficient influence curve satisfies a stronger collaborative double robustness property presented above.

Statistical properties of IPCW-R-TMLE. The above-mentioned robustness property of the estimating function $D(Q^r, g^r, g)$ has immediate implications for the statistical properties of a solution $\Psi(Q_n^{r*})$ such that $\sum_i D(Q_n^{r*}, g_n^r, g_n) = 0$. Firstly, under appropriate regularity conditions, if g_n consistently estimates g_0, then ψ_n will be a consistent and asymptotically linear estimator of ψ_0. In addition, if $g_n(A \mid X)$ and its target $g_0(A \mid X)$ are only functions of the reduced-data structure O^r so that g_n^r / g_n converges to 1, then ψ_n is consistent and asymptotically linear if either Q_n^{r*} consistently estimates Q_0^r, or g_n consistently estimates g_0, and if both estimates are consistent, then the estimator ψ_n is more efficient than an efficient estimator based on n i.i.d.

observations of the reduced-data structure $O^r \sim P^r_{Q^r_0, g^r_0}$ only. In our example with $O^r = (W, A, Y)$, if g_0 is consistently estimated, the IPCW-R-TMLE is asymptotically more efficient and less biased than the R-TMLE (which will be biased if there is time-dependent confounding), and if there is no time-dependent confounding so that $g^r_0/g_0 = 1$ and the estimated weights converge to 1, then the IPCW-R-TMLE is double robust with respect to misspecification of either g^r_0 or Q^r_0, just like the R-TMLE.

A.17 Collaborative Double Robust TMLE

Summary. A TMLE of a causal effect of an intervention requires an estimator of the conditional distribution of an intervention node, given its parents, across all intervention nodes, where this combined set of intervention-node-specific conditional distributions is called the intervention assignment mechanism (such as treatment mechanism). If the estimator of the intervention-assignment mechanism converges to the truth, then the TMLE will be asymptotically unbiased. However, including correct parent nodes for an intervention node that play no role in the g-computation formula for the target parameter only hurts the finite sample bias reduction and can dramatically increase the variance of the TMLE. This suggests that the goal should not be to estimate the true intervention-assignment mechanism but the true conditional distributions of the intervention nodes that condition on sufficient reduction of the true parent nodes so that the desired bias reduction of the TMLE is achieved. The collaborative double robustness of the efficient influence curve and the TMLE formalizes this concept of a sufficient adjustment set for the intervention assignment mechanism, showing that only functions of parent nodes that explain the residual bias of the initial estimator of the g-computation factor of the data-generating distributions need to be included. This collaborative double robustness of the efficient influence curve implies another fundamental invariance property of the TMLE when applied to an infinite sample of the true probability distribution: If the initial estimator is already targeted with a sufficient intervention-assignment mechanism, then the TMLE will not further modify the initial estimator, even when it uses another sufficient intervention-assignment mechanism besides that used by the initial estimator. These fundamental insights yield the theoretical underpinnings of the C-TMLE. The C-TMLE at infinite sample size (i.e., $P_n = P_0$) and its properties are presented.

Let $O = \Phi(C, X) \sim P_0$ for some many-to-one mapping Φ, and consider a CAR censored-data model that assumes some model $\mathcal{Q}$ on the distribution of X, and assumes, minimally, that the conditional distribution g_0 of C, given X, satisfies CAR. Let $\mathcal{G}$ be the model for g_0. Let $\Psi(Q_0)$ be a target parameter for some $Q_0 = Q(P_0)$.

Firstly, we will consider the TMLE algorithm at infinite sample size, so that the empirical probability distribution function P_n is replaced by P_0. In the TMLE, we require that $\frac{d}{d\epsilon}L(Q_g(\epsilon))\Big|_{\epsilon=0} = D(Q,g)$ for some loss function L, fluctuation working model $\{Q_g(\epsilon) : \epsilon\}$ through an initial Q, and estimating function D. As a consequence, if we apply the TMLE to an initial Q using a certain g, then we obtain a solution Q^* (indexed by g used in the working fluctuation model) so that $P_0D(Q^*,g) = 0$. These functions D are chosen such that $P_0D(Q,g_0) = 0$ implies $\Psi(Q) = \Psi(Q_0)$ [or, minimally, are such that $P_0D(Q,g_0) = 0$ if $\Psi(Q) = \psi_0$], even if Q itself is misspecified. In this way, using the true g_0 in the TMLE, we obtain a Q^* with $\Psi(Q^*) = \psi_0$ that has thereby removed all the bias of the initial $\Psi(Q^0)$ with respect to the true target ψ_0. However, the estimating functions we will use satisfy a stronger collaborative robustness property in terms of a specified subset $\mathcal{G}(Q,P_0)$ of the parameter space $\mathcal{G}$ for g_0, which includes the true g_0. If $g \in \mathcal{G}(Q,P_0)$, then

$$P_0D(Q,g) = 0 \text{ implies } \Psi(Q) = \Psi(Q_0).$$

In a coarsening at random censored-data model, this set $\mathcal{G}(Q,P_0)$ includes any true conditional distribution of the censoring variable, conditioning on a reduction of the full data that captures a specified difference defined in terms of $Q - Q_0$ (Appendix A.8). In particular, if Q converges to Q_0, then the set $\mathcal{G}(Q,P_0)$ grows to the set $\mathcal{G}$ of all distributions. In particular, by applying this result at $Q = 0$, it follows that $\mathcal{G}(Q,P_0)$ includes distributions that do not condition on variables used by the true g_0 that Q_0 does not depend on.

Suppose now that the TMLE Q_n^* uses an estimator g_n that converges to a $g_0(Q^*) \in \mathcal{G}(Q^*,P_0)$. In that case, the corresponding TMLE Q_n^* that solves $P_nD(Q_n^*,g_n) = 0$ will asymptotically solve $P_0D^*(Q^*,g_0(Q^*)) = 0$, which implies $\Psi(Q^*) = \psi_0$. That is, the desired asymptotic bias reduction can be obtained by using an estimator g_n that is inconsistent for the true g_0 but that converges to an element $g_0(Q^*) \in \mathcal{G}(Q^*,P_0)$. This suggests that we should be using collaborative estimators g_n in TMLE that aim to converge to such a $g_0(Q^*)$ that takes into account the residual bias of the initial estimator. We state the following theorem laying out two properties of the TMLE algorithm when applied to P_0 (instead of finite data set P_n).

Theorem A.3. *For a given Q and P_0, let $\mathcal{G}(Q,P_0) \subset \mathcal{G}$ be such that $g \to P_0D(Q,g)$ is constant in $\mathcal{G}(Q,P_0)$, and that for each $g \in \mathcal{G}(Q,P_0)$ $P_0D(Q,g) = 0$ implies $\Psi(Q) = \psi_0$. Define $f(\epsilon) = P_0L(Q_g(\epsilon))$ and assume $\frac{d}{d\epsilon}L(Q_g(\epsilon))\Big|_{\epsilon=0} = D(Q,g)$. Assume f has a unique local minimum satisfying $f'(\epsilon) = 0$. For a TMLE Q that used $g^0 \in \mathcal{G}(Q,P_0)$ to fluctuate an initial Q^0, we have $P_0D(Q,g^0) = 0$ and thereby $\Psi(Q) = \psi_0$. Consider a TMLE Q^* that uses this TMLE Q as initial estimator, and uses another $g \in \mathcal{G}(Q,P_0)$. Then $Q^* = Q$, and thus $\Psi(Q^*) = \psi_0$.*

Proof. By the constant property, $P_0D(Q,g) = P_0D(Q,g^0)$, and since $P_0D(Q,g^0) = 0$, we also have that $P_0D(Q,g) = 0$. Recall that the TMLE update will calculate: $\epsilon^1 = \arg\min_\epsilon P_0L(Q_g(\epsilon))$. By being a minimum of $f(\epsilon) \equiv P_0L(Q_g(\epsilon))$ at an interior point, we have that ϵ^1 solves the derivative equation $0 = f'(\epsilon) \equiv \frac{d}{d\epsilon}f(\epsilon)$. By assumption, the derivative $f'(\epsilon)$ has only one solution with $f'(\epsilon) = 0$: For example, the fluctuation $f(\epsilon)$ has only one local maximum. However, $\epsilon = 0$ is a solution since

$f'(0) = P_0 D(Q, g)$, which equals zero since $g \in \mathcal{G}(Q, P_0)$, as shown above. Thus, the TMLE algorithm will set $\epsilon_n^1 = 0$ and thus not update the initial Q. □

This proves that, not only does the TMLE algorithm only require a $g_0(Q) \in \mathcal{G}(Q, P_0)$ in order to achieve the full asymptotic bias reduction, but, in addition, the TMLE algorithm using such a $g_0(Q)$ will not update an initial Q that already solves a $P_0 D(Q, g) = 0$ for a $g \in \mathcal{G}(Q, P_0)$. That is, TMLE is "smart enough" to keep an unbiased initial (TMLE) unbiased. This motivates the following C-TMLE at P_0.

Theorem A.4. *Suppose that we are given a sequence $g^1, \ldots, g^K$ of candidates satisfying the following property: For any Q, there exists a $k \in \{1, \ldots, K\}$ so that $g^k \in \mathcal{G}(Q, P_0)$ (e.g., $g^K = g_0$). Consider the following C-TMLE algorithm. Start with Q^0, g^1; as the first step, compute TMLE Q^{1*} based on initial Q^0 using g^1; as the second step, compute TMLE Q^{2*} based on initial Q^{1*} using g^2, and, in general, at the kth step, compute TMLE Q^{k*} updating Q^{k-1*} using g^k, $k = 1, \ldots, K$. Select $k_0 = \arg\min_k P_0 L(Q^{k*})$, where we select the smallest among the minima. The output of the C-TMLE is now $(Q^* \equiv Q^{k_0*}, g^{k_0*})$ and the corresponding C-TMLE $\Psi(Q^*)$ of ψ_0. Assume that, for each k, if $g^k \in \mathcal{G}(Q^{k*}, P_0)$, then $g^{k+1} \in \mathcal{G}(Q^{k*}, P_0)$.*

Properties. *This procedure generates K TMLEs $(Q^{1^*}, g^1), \ldots, (Q^{K^*}, g^K)$. This sequence of candidate TMLEs has the following properties. (1) There exists a smallest $k_0 \in \{1, \ldots, K\}$ so that $\Psi(Q^{k_0*}) = \psi_0$; (2) for $k \geq k_0$, $Q^{k^*} = Q^{k_0*}$, and, in particular, $\Psi(Q^{k^*}) = \psi_0$; and (3) $P_0 L(Q^{k^*})$ is decreasing in $k \in \{1, \ldots, k_0\}$ and constant for $k \geq k_0$. The C-TMLE selects this smallest k_0 and thus satisfies $\Psi(Q^*) = \psi_0$.*

The existence of a $g_k \in \mathcal{G}(Q, P_0)$ is guaranteed by making $g_K = g_0$. The conservation part of this property can typically be arranged by, for each k, making g^{k+1} a more nonparametric fit of g_0 than g^k. For example, g^{k+1} could be a conditional distribution of C, adjusting for an extra binary variable beyond the k variables that g^k adjusted for. This extra binary variable could be selected from among a set of candidates as the one that yields the maximal decrease in risk for the resulting Q^{k*}, thereby allowing for algorithms that build sequences $(g^k : k)$ that are maximally effective in bias reduction. In this way, the elements g^k also approximate the true g_0 when k increases. We wish to select this smallest k_0 since it corresponds with a TMLE that uses the smallest sufficient approximation of g_0. The additional efforts in bias reduction for steps $k > k_0$ in the C-TMLE algorithm are useless at P_0, and will induce unnecessary variance and bias for finite samples.

In the above C-TMLE algorithm the next TMLE in the sequence used the previous TMLE as initial estimator, thereby guaranteeing that the risk $P_0 L(Q)$ (i.e., the expectation of the loss function) of the candidate TMLEs decreases in k. If, just by virtue of using the next g^{k+1}, the next TMLE $Q^{k+1,*}$ already decreases the risk, i.e., $P_0 L(Q^{k+1,*}) < P_0 L(Q^{k,*})$, when using the same initial estimator as $Q^{k,*}$ uses, then we do not have to update the initial estimator. With this modification of the above C-TMLE algorithm, the updating of the initial estimator, which involves extra fitting of the data, is preserved for when it is necessary. As a consequence, the resulting C-TMLE algorithm can be applied with long sequences $(g^k : k)$ that slowly approximate g_0 in k and only now and then update the initial estimator for the TMLEs.

The empirical counterpart of this algorithm represents the C-TMLE algorithm one applies to a data set. That is, in the above description of C-TMLE, Q plays the role of an initial estimator $\hat{Q}^0(P_n)$, g^k plays role of the kth estimator $\hat{g}^k(P_n)$ of g_0, and P_0 is replaced by P_n. In addition, minimizing the risk $P_0 L(Q^{k*})$ over the candidates indexed by k to select the desired TMLE among the sequence of TMLEs is replaced by minimizing the cross-validated risk of the estimator $P_n \to \hat{Q}^{k*}(P_n)$, so that k_0 is replaced by the optimal cross-validation selector for which oracle results are available.

A.18 Asymptotic Linearity of (C-)TMLE

Summary. We provide a template for proving the asymptotic linearity of the C-TMLE and explain the conditions.

Consider a TMLE or C-TMLE Q_n^* with corresponding g_n, which solves the efficient influence curve estimating equation or some other estimating equation:

$$0 = P_n D^*(Q_n^*, g_n).$$

It is a reasonable assumption that Q_n^* converges to some element Q^* in the model for Q_0, where Q^* is not necessarily equal to the true Q_0. We assume that consistency has been established in the sense that g_n converges to a $g_0 \in \mathcal{G}(Q^*, P_0)$, so that $\Psi(Q^*) = \Psi(Q_0) = \psi_0$. Recall that $\mathcal{G}(Q^*, P_0)$ is such that for each $g \in \mathcal{G}(Q^*, P_0)$ $P_0 D^*(Q^*, g) = 0$ implies $\Psi(Q^*) = \psi_0$. For notational convenience, we will also denote the limit of g_n by g_0 even though it does not need to represent the actual censoring mechanism of the data-generating experiment. Given the consistency of Q_n^* and g_n, we will also have that $P_0 D^*(Q^*, g_0) = 0$.

To derive the influence curve of $\Psi(Q_n^*)$, the asymptotic linearity theorem below assumes that the limit g_0 of the selected censoring mechanism estimator satisfies:

$$P_0 D^*(Q_n^*, g_0) = \psi_0 - \Psi(Q_n^*) + (P_n - P_0) IC_Q + o_P(1/\sqrt{n}) \tag{A.11}$$

for some $IC_Q \in L_0^2(P_0)$. The $o_P(1/\sqrt{n})$ can be replaced by $O\left(\left\|\Psi(Q_n^*) - \psi_0\right\|^2\right)$ as well. In the special case where $IC_Q = 0$, the influence curve does not involve a contribution requiring the analysis of a function of Q_n^*. This potential important simplification of the influence curve allows straightforward calculation of standard errors for the C-TMLE. This assumption is best illustrated with an example, which we provide after the theorem below.

Theorem A.5. *Let $(Q, g) \to D^*(Q, g)$ be a well-defined function that maps any possible (Q, g) into a function of O. Let $O_1, \ldots, O_n \sim P_0$ be i.i.d. and let P_n be the empirical probability distribution. Let $Q \to \Psi(Q)$ be a d-dimensional parameter, where $\psi_0 = \Psi(Q_0)$ is the parameter value of interest. In the following template*

for proving the asymptotic linearity of $\Psi(Q_n^*)$ as an estimator of $\Psi(Q_0)$, Q_n^* and g_n represent a (C-)TMLE of Q_0, coupled with an estimator g_n used in the TMLE step, but it can be any estimator. Let Q^* and g_0 denote the limits of Q_n^* and g_n. Make the following assumptions.

Efficient influence curve estimating equation. $0 = P_n D^*(Q_n^*, g_n)$.

Censoring mechanism estimator is nonparametric *enough*. $P_0 D^*(Q^*, g_0) = 0$ *and* $\Psi(Q^*) = \psi_0$.

Consistent estimation of D^*. $P_0(D^*(Q_n^*, g_n) - D^*(Q^*, g_0))^2 \to 0$ *in probability, as* $n \to \infty$. *The same is assumed if one component of* (Q_n^*, g_n) *is replaced by its limit* (Q^*, g_0).

Donsker class. $\{D^*(Q, g) : Q, g\}$ *is* P_0*-Donsker, where* (Q, g) *vary over sets that contain* (Q_n^*, g_n), (Q^*, g_n), (Q_n^*, g) *with probability tending to 1.*

Asymptotic linearity condition for censoring mechanism estimator. *Define the mapping* $g \to \Phi(g) \equiv P_0 D^*(Q^*, g)$. *Assume* $\Phi(g_n) - \Phi(g_0) = (P_n - P_0) IC_{g_0} + o_P(1/\sqrt{n})$ *for some mean-zero function* $IC_{g_0} \in L_0^2(P_0)$.

Asymptotic linearity of Q_0-estimator.

$$P_0 D^*(Q_n^*, g_0) = \psi_0 - \Psi(Q_n^*) + (P_n - P_0) IC_{Q^*} + o_P(1/\sqrt{n}). \tag{A.12}$$

Second-order term. *Define also the second-order term*

$$R_n = P_0\{D^*(Q_n^*, g_n) - D^*(Q_n^*, g_0)\} - P_0\{D^*(Q^*, g_n) - D^*(Q^*, g_0)\},$$

and assume $R_n = o_P(1/\sqrt{n})$. *Note* R_n *is a second-order term involving the product of the differences* $Q_n^* - Q^*$ *and* $g_n - g_0$.

Then ψ_n is an asymptotically linear estimator of ψ_0 at P_0 with the influence curve

$$IC(P_0) = D^*(Q^*, g_0, \psi_0) + IC_{Q^*} + IC_{g_0}.$$

In particular, $\sqrt{n}(\psi_n - \psi_0)$ converges in distribution to a multivariate normal distribution with mean zero and covariance matrix $\Sigma_0 = E_0 IC(P_0) IC(P_0)^\top$.

Proof. The principal equations are $0 = P_n D^*(Q_n^*, g_n) = P_0 D^*(Q^*, g_0)$, and the first second-order-term condition $P_0 D^*(Q_n^*, g_0) = \psi_0 - \Psi(Q_n^*) + (P_n - P_0) IC_{Q^*} + o_P(1/\sqrt{n})$. This yields

$$\Psi(Q_n^*) - \psi_0 = (P_n - P_0)\{D^*(Q_n^*, g_n) + IC_{Q^*}\} + P_0\{D^*(Q_n^*, g_n) - D^*(Q_n^*, g_0)\} + o_P(1/\sqrt{n}).$$

By the consistency condition and Donsker condition, the first term on the right-hand side equals $(P_n - P_0) D^*(Q^*, g_0) + o_P(1/\sqrt{n})$ (van der Vaart and Wellner 1996). The second term on the right-hand-side equals R_n plus the term $\Phi(g_n) - \Phi(g) = P_0\{D^*(Q^*, g_n) - D^*(Q^*, g_0)\}$. The asymptotic linearity condition on the censoring mechanism estimator shows that this equals $(P_n - P_0) IC_{g_0} + o_P(1/\sqrt{n})$. This completes the proof. □

Illustration of condition (A.12). Suppose $O = (W, A, Y) \sim P_0$, the model for P_0 is nonparametric, and the target parameter is $\psi_0 = E_0[E_0(Y \mid A = 1, W)]$. Suppose g_n converges to some true conditional distribution of A, given W^s, for some reduction W^s of W, and we will denote the latter by g_0. For any g_0, we have

$$\begin{aligned} P_0 D^*(Q_n^*, g_0) &= P_0 \frac{A}{g_0(A \mid W^s)}(\bar{Q}_0(1, W) - \bar{Q}_n^*(1, W)) + \bar{Q}_n^*(1, W) - \Psi(Q_n^*) \\ &= P_0 \left\{ \frac{A}{g_0(A \mid W^s)} - 1 \right\} (\bar{Q}_0(1, W) - \bar{Q}_n^*(1, W)) + \psi_0 - \Psi(Q_n^*). \end{aligned}$$

Verification of condition (A.12) requires showing that the first term involving the expectation with respect to P_0 is asymptotically linear with some influence curve IC_{Q^*}. Firstly, consider the case that g_0 is the true conditional distribution of A, given W, i.e., $W^s = W$. In that case, by conditioning on W, and noting that $E_0(A/g_0(A \mid W) - 1) = 0$, it follows that this term equals zero, so that (A.12) holds with $IC_{Q^*} = 0$. Secondly, consider the case where $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$ only depends on W through W^s, and that $\bar{Q}_n^*$ is only a function of W^s. In this case, the residual bias $\bar{Q}_0(1, W) - \bar{Q}_n^*(1, W)$ is only a function of W^s. As a consequence, by conditioning on W^s, it follows again that this first term equals zero, so that (A.12) holds with $IC_{Q^*} = 0$. However, if we only know that $\bar{Q}_n^*$ converges to a $\bar{Q}^*$ for which the asymptotic residual bias $\bar{Q}^*(1, W) - \bar{Q}_0(1, W)$ is only a function of W^s, then this first term equals $P_0 \left\{ \frac{A}{g_0(A|W^s)} - 1 \right\} (\bar{Q}^* - \bar{Q}_n^*)(1, W)$, which might potentially contribute an influence curve term IC_{Q^*}. The latter term would require showing that this integrated difference $\bar{Q}_n^* - \bar{Q}^*$ is asymptotically linear. In practice, one might consider adjusting for $\bar{Q}_n^0$ or, using an iterative procedure, for $\bar{Q}_n^*$, in g_n, so that W^s will include this potential dependence of $\bar{Q}_n^*$ on covariates that do not theoretically affect Y.

To summarize, (1) if g_n converges to a true conditional distribution $g_0(A \mid W^s)$ that conditions minimally on all relevant confounders (i.e., all variables that the conditional mean of Y depends on), and the estimator $\bar{Q}_n^*$ is a function of W^s only with probability tending to 1, then it follows that condition (A.12) holds with $IC_{Q^*} = 0$; (2) if, on the other hand, g_n is a collaborative estimator that converges to the true conditional distribution $g_0(A \mid W^s)$ that conditions on a rich enough reduction W^s of W that captures the asymptotic residual bias $\bar{Q}^*(1, W) - \bar{Q}_0(1, W)$ (which is sufficient for the consistency of the C-TMLE), then the estimator $\bar{Q}_n^*$ will contribute an IC_{Q^*} to the influence curve of $\Psi(Q_n^*)$ through $P_0 \left\{ \frac{A}{g_0(A|W^s)} - 1 \right\} (\bar{Q}^* - \bar{Q}_n^*)(1, W) \approx (P_n - P_0) IC_{Q^*}$.

A.19 Efficiency Maximization and TMLE

Summary. Consider estimating a pathwise differentiable parameter on a semiparametric model based on n i.i.d. observations. The TMLE is a consistent, asymptotically linear, locally efficient substitution estimator of the target pa-

rameter under appropriate regularity conditions. The asymptotic efficiency corresponds with asymptotic optimal estimation [usually implying remarkable robustness such as double robustness in censored-data models that satisfy the CAR assumption, van der Laan and Robins (2003)], while being a substitution estimator guarantees that the estimator respects the global constraints on the target parameter imposed by the statistical model and the target parameter mapping. The latter allows the estimator to be robust under sparsity. Another property of interest of an estimator is that it is guaranteed to asymptotically outperform a user supplied class of asymptotically linear estimators, i.e., even when it is not asymptotically efficient, but still asymptotically linear, it will outperform each of the estimators in this class. That is, the estimator is guaranteed to asymptotically dominate a certain user-supplied class of asymptotically linear estimators. This can be achieved with empirical efficiency maximization (EEM) as introduced in Rubin and van der Laan (2008) for empirical efficiency over parametric models, refined by Tan (2008) to preserve double robustness, and presented in terms of cross-validation to select among candidate C-TMLE in van der Laan and Gruber (2010). In the next sections we demonstrate in great generality how EEM and TMLE can be combined into a TMLE that also satisfies this dominance property. It involves an application of loss-based super learning with the squared-efficient-influence-curve loss function and a library of candidate TMLEs. For RCTs it guarantees that the resulting TMLE dominates a user-supplied class of asymptotically linear estimators.

Super learner with squared efficient influence curve loss. Let $O \sim P_0 \in \mathcal{M}$, and let $\Psi : \mathcal{M} \to \mathbb{R}$ be the target parameter of interest. Let $O_1, \ldots, O_n$ be i.i.d. copies of O. Suppose that $\Psi(P_0)$ only depends on P_0 through a parameter Q_0. We will also use the notation $\Psi(Q_0)$. Let L be a loss function for Q_0 so that $Q_0 = \arg\min_{Q \in \mathcal{Q}} P_0 L(Q)$. Let $D^*(P)$ be the efficient influence curve at P of the parameter $\Psi : \mathcal{M} \to \mathbb{R}$, and suppose that it depends on $Q(P)$ and $g(P)$ for some other (nuisance) parameter g. In addition, for given values Q, g, let $\{Q_g(\epsilon) : \epsilon\} \subset \mathcal{M}$ be a submodel with $Q_g(\epsilon = 0) = Q$ satisfying $D^*(Q, g) \in \langle \frac{d}{d\epsilon} L(Q_g(\epsilon))\big|_{\epsilon=0} \rangle$. A TMLE can now be defined in terms of an initial estimator $Q_n^0 = \hat{Q}^0(P_n)$ of Q_0, an estimator $g_n = \hat{g}(P_n)$ of g_0, and an iterative TMLE-updating algorithm resulting in a TMLE $Q_n^* = \hat{Q}^*(P_n)$ solving $P_n D^*(Q_n^*, g_n) = 0$.

Consider a collection of initial estimators $\hat{Q}_j : \mathcal{M}_{NP} \to \mathcal{Q}$, $j = 1, \ldots, J$, of Q_0. This provides us with a collection of candidate TMLEs $\hat{Q}_j^*$, $j = 1, \ldots, J$. Let $\hat{Q}_\alpha = f_\alpha(\hat{Q}_j : j)$ be a combination of the J initial estimators indexed by a weight-vector α. For example, $\hat{Q}_\alpha = \sum_j \alpha(j)\hat{Q}_j$. Note that $\hat{Q}_\alpha$ is just another initial estimator indexed by a vector of weights α. This family of candidate initial estimators $\hat{Q}_\alpha$ indexed by a choice α includes the discrete choices $\hat{Q}_j$, $j = 1, \ldots, J$. This family of candidate initial estimators $\hat{Q}_\alpha$ generates a corresponding family of TMLEs given by $\hat{Q}_\alpha^*$ indexed by α. We wish to select among these candidate TMLEs. (The method

below also applies for selection among candidate C-TMLEs $\hat{Q}^*_\alpha$ involving a collaborative estimator $g_{n,\alpha}$ of g_0.) For that purpose, we need a loss function for Q_0 so that we can use the cross-validation selector. We wish to choose a loss function that selects the estimator with the best asymptotic efficiency among all the α-specific candidate TMLEs of ψ_0. A related goal (and equivalent goal if g_0 is known) is to choose a loss function that selects the estimator $\hat{Q}^*_\alpha$ that yields the best estimator $D^*(\hat{Q}^*_\alpha, g_0)$ of the true efficient influence curve $D^*(Q_0, g_0)$. We demonstrate how both goals can be achieved.

This is a sensible goal if one believes that all candidate TMLEs are considered asymptotically linear estimators of ψ_0. We will first consider the case where we have available a consistent estimator g_n of g_0, and, in this case, we wish to make sure that the proposed selector achieves its goal. For example, g_0 might be known, such as in an RCT, or the design provides enough knowledge about g_0 (e.g., it is known that censoring is independent) such that a good consistent estimator of g_0 will be available. Either way, g_0 is typically a much easier to estimate parameter than Q_0, so that utilizing an estimator of g_0 in order to improve the estimation of Q_0 is sensible.

We could now apply loss-based super learning, with this library of candidate estimators $\hat{Q}^*_\alpha$ indexed by α, to estimate Q_0 with the following targeted loss function:

$$L_{g_0}(Q) = \{D^*(Q, g_0)\}^2.$$

Since our candidate estimators are supposedly consistent for ψ_0, this is a valid loss function if $P_0\{D^*(Q, g_0)\}^2$ is minimized at $Q = Q_0$ among all Qs with $\Psi(Q) = \psi_0$. We now explain why this is indeed a valid loss function. The basic point is that as long as Q correctly specifies ψ_0 (and in some models one needs to correctly specify a larger parameter of Q_0), then $D^*(Q, g_0)$ is typically a gradient of the target parameter mapping $\Psi : \mathcal{M}(g_0) \to \mathbb{R}$ for the model $\mathcal{M}(g_0) \subset \mathcal{M}$ where g_0 is known. As a consequence, $D^*(Q_0, g_0) = \Pi(D^*(Q, g_0) \mid T_{Q_0}(P_0))$, where $T_{Q_0}(P_0)$ is the tangent space of model $\mathcal{M}(g_0)$ and Π is the projection operator in the Hilbert space $L^2_0(P_0)$. This proves that $\| D^*(Q_0, g_0) \|^2_{P_0} \leq \| D^*(Q, g_0) \|^2_{P_0}$. More importantly, by the theorem of Pythagoras, this proves that for a Q that correctly specifies the desired part of Q_0 (including ψ_0), we have

$$\| D^*(Q, g_0 \|^2_{P_0} - \| D^*(Q_0, g_0) \|^2_{P_0} = \| D^*(Q, g_0) - D^*(g_0, Q_0) \|^2 .$$

The left-hand side equals the loss-based dissimilarity $P_0\{L_{g_0}(Q) - L_{g_0}(Q_0)\}$. This proves that indeed the loss function L_{g_0} is a valid loss function with a loss-based dissimilarity equal to a squared $L^2(P_0)$-norm of $D^*(Q, g_0) - D^*(Q_0, g_0)$!

This argument relies on $D^*(Q, g_0)$ being a gradient in the model where g_0 is known and Q correctly specifies ψ_0. This can be further supported as follows. By Theorem 1.3 in van der Laan and Robins (2003) for CAR censored-data models for $O = \Phi(C, X) \sim P_0$ with the censoring mechanism $g_0(C \mid X)$ being known, a class of gradients of the pathwise derivative can typically be represented as $D^*(Q, g_0) = D_{IPCW}(\Psi(Q), \theta(Q), g_0) + D_{CAR}(Q, g_0)$ for any Q satisfying $\Psi(Q) = \psi_0$ and $\theta(Q) = \theta_0$, where in many cases the additional nuisance parameter $\theta_0 = \theta(Q_0)$

is not present. Here ψ_0, θ_0 represent the part of Q_0 that need to be consistently estimated, while the remaining part of Q_0 is protected against misspecification in the sense that $P_0 D^*(Q, g_0) = 0$ as long as $\Psi(Q) = \psi_0$ and $\theta(Q) = \theta_0$. Here D_{IPCW} is an IPCW estimating function and $D_{CAR}(Q, g_0) \in T_{CAR}(P_0)$ is an element in the tangent space $T_{CAR}(P_0) = \{V(O) : E_{g_0}(V(O) \mid X) = 0\}$ of g_0 when only assuming CAR on g_0. The optimal choice in this set of gradients is achieved at $Q = Q_0$ so that $D^*(Q_0, g_0) = D_{IPCW}(\psi_0, \theta_0, g_0) + D_{CAR}(Q_0, g_0)$. This shows that $Q \to P_0\{D^*(Q, g_0)\}^2$ is minimized at Q_0 over all Q with $\Psi(Q) = \psi_0$ and $\theta(Q) = \theta_0$. As a consequence, indeed, $\{D^*(Q, g_0)\}^2$ is a valid loss function to select Q_0 among a class of Q with $\Psi(Q) = \psi_0$ and $\theta(Q) = \theta_0$. For the sake of presentation (and the examples covered in this book do not have a θ_0 due to our observed data models being nonparametric), we consider the case that θ_0 is not present. In particular, if g_0 is known, then the TMLE $\Psi(Q_n^*)$ using the known g_0 is asymptotically linear with influence curve $D^*(Q, g_0)$ with Q being the limit of Q_n^*, so that the optimal influence curve among all these influence curves is the efficient influence curve $D^*(Q_0, g_0)$. In this special case where g_0 is known, the cross-validation selector based on $L_{g_0}(Q)$ corresponds with minimizing the variance of the influence curves of the candidate TMLEs $\hat{Q}_\alpha^*$. We conclude that $Q_0 = \arg\min_Q P_0 L_{g_0}(Q)$, where the minimum is taken over all $Q \in \mathcal{Q}$ with $\Psi(Q) = \psi_0$.

Given a cross-validation scheme $B_n \in \{0, 1\}^n$ with corresponding empirical distributions P^1_{n,B_n}, P^0_{n,B_n} for the B_n-specific validation and training sample, we select the TMLE indexed by

$$\alpha_n = \arg\min_\alpha E_{B_n} P^1_{n,B_n} L_{g_0}(\hat{Q}^*_\alpha(P^0_{n,B_n})).$$

The resulting estimator of ψ_0 is given by

$$\psi_n^* = \Psi(\hat{Q}^*_{\alpha_n}(P_n)).$$

The cross-validation selector needs to be applied to estimators that are consistent for ψ_0 at a faster rate than the rate at which Q_0 can be estimated with respect to the loss-based dissimilarity. For example, if all the candidate estimators $\Psi(\hat{Q}^*_\alpha(P_n))$ are asymptotically linear, then this holds. It is also possible to use the above loss function by plugging in a separate estimator for ψ_0 in a representation $D^*(Q_0, g_0, \psi_0)$ of the efficient influence curve, so that both g_0 and ψ_0 are treated as nuisance parameters of this loss function for Q_0 that need to be estimated once and for all before the selection process starts. In many examples of interest, $D^*(Q, g_0) = D(Q, g_0) - \psi_0$ for some $D(Q, g_0)$, in which case, we can define the loss as

$$L_{g_0}(Q) = D(Q, g_0)^2,$$

which no longer depends on ψ_0. The latter is now a valid loss function over all Q (i.e., no need to restrict to Q with $\Psi(Q) = \psi_0$).

Before we proceed in our discussion of the theoretical properties of this cross-validation selector α_n in the next section, we conclude this section with a few remarks. Firstly, one could decide not to cross-validate the candidate TMLEs, such

that

$$\alpha_n^e = \arg\min_\alpha P_n L_{g_0}(\hat{Q}_\alpha^*(P_n)).$$

This includes the case where $\hat{Q}_\alpha(P_n)$ is constant in P_n so that $\{\hat{Q}_\alpha : \alpha\}$ represents a parametric model, and $\hat{Q}_\alpha^*$ represents the TMLE that updates this particular non-random initial $\hat{Q}_\alpha$. In this special case, the empirical α_n^e corresponds with empirical efficiency as defined in Rubin and van der Laan (2008) for computing the optimal parameter value of a parametric model that maximizes empirical efficiency of the resulting double robust estimator. Even though α_n^e is appropriate for parametric models, we strongly recommend the cross-validation selector α_n when $\hat{Q}_\alpha$ are adaptive estimators. Our oracle result below for the cross-validation selector α_n proves that α_n will be robust against adaptive initial estimators.

We also note that in great generality $D^*(Q, g_0, \psi_0)$ is linear in Q. For a smooth parametric family $\{Q_\alpha : \alpha\}$, this linearity makes $D^*(Q_\alpha, g_0, \psi_0)^2$ a nice smooth function in α, so that the computation of α_n or α_n^e is computationally tractable. For example, if $Q_\alpha = \sum_j \alpha_j Q_j$, then $D^*(Q_\alpha, g_0, \psi_0) = \sum_j \alpha_j D^*(Q_j, g_0, \psi_0)$, so that optimizing $\alpha \to P_0\{D^*(Q_\alpha, g_0, \psi_0)\}^2$ is equivalent with linear least squares regression, which can be done with simple standard software.

We remark that one could also select among candidate (C)-TMLEs $\hat{Q}_\alpha^*$ by minimizing over α an estimator of the variance of $\Psi(\hat{Q}_\alpha^*(P_n))$. That is, if this estimator has influence curve $IC_\alpha(P_0)$, then we could estimate its variance with $E_{B_n} P^1_{n,B_n} IC^2_{\alpha,P^0_{n,B_n}}/n$, where $IC_{\alpha,P^0_{n,B_n}}$ is an estimator of the influence curve $IC_\alpha(P_0)$ based on the training sample P^0_{n,B_n} only. This is slightly different from the above selector α_n since the influence curve of the TMLE $\Psi(\hat{Q}_\alpha^*(P_n))$ equals $D^*(Q_\alpha^*, g_0)$ plus a term due to estimating g_0 with g_n. The latter contribution improves the influence curve relative to using the true g_0. As shown in van der Laan and Robins (2003) the influence curve of $\Psi(\hat{Q}_\alpha^*(P_n))$ can still be represented as $IC_\alpha(P_0) = D_{IPCW}(\psi_0, g_0) + D_{CAR}(Q_\alpha^*, g_0)$, where the element $D_{CAR}(Q_\alpha^*, g_0) \in T_{CAR}(P_0)$ is a sum of the element in $T_{CAR}(P_0)$ it would have been for known g_0 plus another term due to estimation of g_0. As a consequence, minimizing the variance of the influence curve $IC_\alpha(P_0)$ over choices Q_α^* (all satisfying $\Psi(Q_\alpha^*) = \psi_0$) can still be represented as minimizing $P_0 L_{g_0,1}(Q_\alpha^*)$ for a valid loss function $L_{g_0,1}(Q_\alpha^*)$ that equals the square of the influence curve of $IC_\alpha(P_0)$. However, the form of the loss-based dissimilarity of L_{g_0}, as established in the next section, shows that the cross-validation selector α_n drives the estimated efficient influence curve to the actual efficient influence curve $D^*(Q_0, g_0)$, even when g_0 is estimated with a consistent estimator g_n. This suggests that, in practice, if g_0 is estimated, the variance of the influence curve of the cross-validated selected estimator $\Psi(\hat{Q}_{\alpha_n}^*(P_n))$ will still closely approximate the choice that would select the influence curve with the smallest variance. Therefore, we suggest that for practical purposes no modification of the loss function $L_{g_0}(Q)$ is necessary when the TMLE uses an estimator g_n of g_0.

Our proposed cross-validation selector based on the square efficient influence curve $L_{g_0}(Q)$ does rely on the availability of a consistent estimator g_n of g_0. If one does not want to rely on the consistency of g_n as an estimator of g_0 (e.g., g_0 has

similar complexity to Q_0), and it is not possible to find a robust version of the loss function L_{g_0} so that L_g remains a valid loss function for Q_0 at misspecified g (as we demonstrate in the example in a later section), then one might still utilize this loss function L_{g_0} that targets the variance, by adding it as a penalty to a loss function L for Q_0 (that is not affected by g_0). For some positive constant a, b, let

$$L_{1,g_0}(Q) = aL(Q) + b\frac{L_{g_0}(Q)}{n}.$$

In this case, even if g_0 is misspecified, the loss function $L_{1,g_0}(Q)$ remains valid. Thus, now (say $a = b = 1$)

$$\alpha_n = \arg\min_\alpha E_{B_n} P^1_{n,B_n} \left\{ L(\hat{Q}^*_\alpha(P^0_{n,B_n})) + L_{g_n}(\hat{Q}^*_\alpha(P^0_{n,B_n})) \right\}.$$

This type of valid targeted loss function was utilized in van der Laan and Gruber (2010) to build C-TMLE and to select among candidate TMLEs, where one should note that $P_0 L_{g_0}(Q)/n$ equals the asymptotic variance of the TMLE $\Psi(Q^*_n)$ of ψ_0, if $g_n = g_0$ and Q denotes the limit of Q^*_n. The first loss function $L(Q)$ drives the selection towards Q_0, regardless of the estimator g_n, while the second loss L_{g_0}/n targets the selection toward minimizing the variance of the resulting TMLE of ψ_0. Such a robust targeted loss function can also be used to select among candidate C-TMLEs $\hat{Q}^*_\alpha$, involving a collaborative estimation procedure of g_0 (van der Laan and Gruber 2010).

A.20 Oracle Inequality of Cross-Validation Selector

Let us now present the oracle inequality for this cross-validation selector α_n, as presented originally in van der Laan and Dudoit (2003). Let $d_{g_0}(Q, Q_0) = P_0\{L_{g_0}(Q) - L_{g_0}(Q_0)\}$ denote the loss-function based dissimilarity. Assume that the loss function is bounded: $M_1 \equiv \sup_Q \mid L_{g_0}(Q) - L_{g_0}(Q_0) \mid < \infty$. In addition, we assume that

$$P_0 \left\{ L_{g_0}(Q) - L_{g_0}(Q_0) \right\}^2 \leq M_2 P_0\{L_{g_0}(Q) - L_{g_0}(Q_0)\}.$$

As explained in van der Laan and Dudoit (2003), the latter assumption corresponds with the loss-based dissimilarity being quadratic in the difference between Q and Q_0. Below, we show that indeed, in great generality, $P_0 L_{g_0}(Q) - P_0 L_{g_0}(Q_0) \leq P_0\{D^*(Q, g_0) - D^*(Q_0, g_0)\}^2$. Thus, to prove the second property of the loss function L_{g_0}, it remains to show that $P_0\{D^{*2}(Q, g_0) - D^{*2}(Q_0, g_0)\}^2 \leq M_2 P_0\{D^*(Q, g_0) - D^*(Q_0, g_0)\}^2$ for some $M_2 < \infty$. The latter trivially holds for bounded D^*:

$$\begin{aligned} &P_0\{D^{*2}(Q, g_0) - D^{*2}(Q_0, g_0)\}^2 \\ &\quad = P_0\{D^*(Q, g_0) - D^*(Q_0, g_0)\}^2 \{D^*(Q, g_0) + D^*(Q_0, g_0)\}^2 \\ &\quad \leq \sup_o \mid \{D^*(Q, g_0) + D^*(Q_0, g_0)\}^2 \mid (o) P_0\{D^*(Q, g_0) - D^*(Q_0, g_0)\}^2, \end{aligned}$$

which completes the proof. This allows us to apply the oracle inequality for the cross-validation selector as presented in van der Laan and Dudoit (2003) providing us with the following result. If the cross-validation selector α_n is defined as a minimizer over a grid with $K(n)$ α-values, then for any $\delta > 0$,

$$Ed_{g_0}(\hat{Q}_{\alpha_n}(P^0_{n,B_n}), Q_0) \leq (1 + 2\delta)E \min_{\alpha} E_{B_n} d_{g_0}(\hat{Q}_{\alpha}(P^0_{n,B_n}), Q_0) + C(M_1, M_2, \delta)\frac{\log K(n)}{n},$$

where $C(M_1, M_2, \delta)$ is a specified constant. The $\tilde{\alpha}_n$ that attains the minimum on the right-hand side is referred to as the oracle selector that selects the α that minimizes the dissimilarity with Q_0 for the given sample P_n. By choosing a grid with width $1/n$, we obtain a grid such that no precision is lost. In that case, the $\log K(n)$ is bounded by a constant times $\log n$. Theorem 1 of van der Laan and Dudoit (2003) also present a finite sample oracle inequality for the case where g_0 in the loss function L_{g_0} is estimated with g_n. From this finite sample inequality it follows that, if g_n converges faster to g_0 than $Q^*_{\alpha_n,n}$ converges to Q_0, then the finite sample oracle inequality is asymptotically equivalent to the above inequality (i.e., the estimation of g_n has an asymptotically negligible effect).

A.21 Loss-Based Dissimilarity

We now want to understand the loss-based dissimilarity $d_{g_0}(Q, Q_0) = P_0\{L_{g_0}(Q) - L_{g_0}(Q_0)\}$ implied by this loss function, so that the oracle result for the cross-validation selector can be interpreted accordingly. Above, we showed that this loss-based dissimilarity is the $L^2_0(P_0)$-norm of $D^*(Q, g_0) - D^*(Q_0, g_0)$, but we provide some additional detail here. Suppose $\Psi(Q) = \psi_0$. As remarked earlier, by Theorem 1.3 in van der Laan and Robins (2003) for CAR censored-data models for the observed data structure $O = \Phi(C, X)$ for some mapping Φ, full-data structure X, and censoring variable C, it follows that $D^*(Q, g_0) = D(\psi_0, g_0) - D_{CAR}(Q, g_0)$, where $D_{CAR}(Q, g_0)$ is an element of $T_{CAR}(P_0)$, $T_{CAR}(P_0) = \{V \in L^2_0(P_0) : E_{g_0}(V(O) \mid X) = 0\}$ is the tangent space of the conditional distribution g_0 of C, given X, when only assuming CAR, $D(\psi_0, g_0)$ is an IPCW function (i.e., a gradient in the model in which g_0 is known), and $D_{CAR}(Q_0, g_0)$ is the projection of $D(\psi_0, g_0)$ onto $T_{CAR}(P_0)$ in $L^2_0(P_0)$. Thus, $D^*(Q_0, g_0) = D(\psi_0, g_0) - \Pi(D(\psi_0, g_0) \mid T_{CAR}(P_0))$, while $D^*(Q, g_0) = D(\psi_0, g_0) - D_{CAR}(Q, g_0)$ with $D_{CAR}(Q, g_0) \in T_{CAR}(P_0)$. Recall $L_{g_0}(Q) = D^{*2}(Q, g_0)$. The risk $P_0 L_{g_0}(Q)$ equals the variance of $D^{*2}(Q, g_0)$ and can be denoted by $\| D^*(Q, g_0) \|^2$, where $\| \cdot \|$ is the inner-product norm in $L^2_0(P_0)$. By the theorem of Pythagoras, we have that

$$\begin{aligned} P_0 L_{g_0}(Q) - P_0 L_{g_0}(Q_0) &= \| D^*(Q, g_0) \|^2 - \| D^*(Q_0, g_0) \|^2 \\ &= \| D_{CAR}(Q, g_0) - D_{CAR}(Q_0, g_0) \|^2 \\ &= \| D^*(Q, g_0) - D^*(Q_0, g_0) \|^2 . \end{aligned}$$

This shows that L_{g_0} is a valid loss function for Q_0 and that its loss-based dissimilarity is a quadratic dissimilarity between Q and Q_0. Moreover, it shows that the loss-based dissimilarity is a direct $L^2(P_0)$ distance between the candidate efficient influence curve $D^*(Q, g_0)$ and the efficient influence curve $D^*(Q_0, g_0)$, or equivalently, between $D_{CAR}(Q, g_0)$ and $D_{CAR}(Q_0, g_0)$.

Thus, the oracle selector $\tilde{\alpha}_n$ selects, for the given sample $O_1, \ldots, O_n$, the estimator among $\{Q^*_\alpha : \alpha\}$ that yields the best estimator of the efficient influence curve $D^*(Q_0, g_0)$ with respect to the $L^2(P_0)$-norm. Since the finite sample and asymptotic behavior of a TMLE is driven by how well the efficient influence curve is estimated (see our asymptotic linearity theorem), this is essentially the best possible (i.e., most targeted) dissimilarity measure, and thereby loss function, for selecting among the candidate TMLEs.

In particular, this result teaches us that, if g_0 is known, the selected TMLE $\Psi(\hat{Q}^*_{\alpha_n}(P_n))$ will be asymptotically at least as efficient as any of the TMLEs $\Psi(\hat{Q}^*_\alpha(P_n))$, and, in case there are several candidate TMLEs that are asymptotically efficient, it is expected to achieve the efficiency bound at a faster rate in sample size than other asymptotically efficient candidate TMLEs. In addition, even if g_0 is estimated and the estimator g_n approaches g_0 faster than Q_n approaches Q_0, the selected TMLE $\Psi(\hat{Q}^*_{\alpha_n}(P_n))$ will be asymptotically equivalent to the oracle selected TMLE $\Psi(\hat{Q}^*_{\tilde{\alpha}_n}(P_n))$ (where the oracle uses the true g_0!), and thereby will yield the best selection with respect to the approximation of the true efficient influence curve $D^*(Q_0, g_0)$. As mentioned earlier, if g_0 is estimated, the selector α_n is not directly concerned with selecting the α-specific TMLE of ψ_0 whose influence curve is optimal, since it ignores that the true influence curve of the α-specific TMLE involves a possible contribution due to estimating g_n (where this contribution equals zero if $\hat{Q}_\alpha$ is consistent for Q_0). However, the oracle inequality shows that, indirectly, it will still get very close to minimizing the actual asymptotic variance.

Since, typically, an element $D_{CAR}(Q, g_0)$ factorizes as $D_{CAR}(Q, g_0) - D_{CAR}(Q_0, g_0) = H_{1,g_0} H_{2,Q-Q_0}$, the loss-based dissimilarity can be represented as a weighted L^2-norm, $P_0 H^2_{1,g_0} H^2_{2,Q-Q_0}$ (which is also a valid norm at misspecified g_0!). The latter also suggests that it might be possible to find an alternative weighted-squared-error-type loss function with the same or similar dissimilarity so that it remains a valid loss function for Q_0 at misspecified g. Such a loss function preserves the double robustness of the resulting TMLE $\Psi(\hat{Q}^*_{\alpha_n}(P_n))$. Indeed, as in the example below, it appears that this is sometimes possible.

A.22 Examples: Loss-Based Dissimilarity

Let us consider an example to demonstrate this last point. Consider the missing-data example $O = (W, \Delta, \Delta Y) \sim P_0$, a nonparametric statistical model, and target parameter $\psi_0 = E_0 Y$. Let $g_0(\delta \mid W) = P_0(\Delta = \delta \mid W)$ and $\bar{Q}_0(W) = E_0(Y \mid W, \Delta = 1)$. In this case, $D^*(Q_0, g_0) = D_{IPCW}(g_0, \psi_0) - D_{CAR}(\bar{Q}_0, g_0)$, where $D_{IPCW}(g_0, \psi_0) = Y\Delta / g_0(1 \mid W) - \psi_0$ and $D_{CAR}(\bar{Q}_0, g_0) = \bar{Q}_0(W)\left(\frac{\Delta}{g_0(1|W)} - 1\right)$, so that by our general result

$$P_0 L_{g_0}(Q) - P_0 L_{g_0}(Q_0) = E_0(\bar{Q} - \bar{Q}_0)^2 \frac{g_0(0 \mid W)}{g_0(1 \mid W)}$$
$$= E_0(Y - \bar{Q}(W))^2 \frac{g_0(0 \mid W)}{g_0(1 \mid W)} - E_0(Y - \bar{Q}_0(W))^2 \frac{g_0(0 \mid W)}{g_0(1 \mid W)}.$$

This shows that the loss function L_{g_0} has a dissimilarity that is equivalent to the dissimilarity implied by the weighted-least-squares *full-data* loss function

$$L_{2,g_0}(\bar{Q}) = (Y - \bar{Q}(W))^2 \frac{g_0(0 \mid W)}{g_0(1 \mid W)}.$$

This full-data loss function could be mapped into an observed-data IPCW version of L_{2,g_0}:

$$L_{IPCW,g_0}(\bar{Q}) = (Y - \bar{Q}_0(W))^2 \Delta \frac{g_0(0 \mid W)}{g_0^2(1 \mid W)}.$$

Note that this loss function $L_{IPCW,g_0}(\bar{Q})$ has the same risk, and thereby loss-based dissimilarity, as L_{g_0}. However, this IPCW loss function has the property that it remains a valid loss function if g_0 is misspecified, so that the resulting TMLE $\Psi(Q^*_{\alpha_n,n})$ remains double robust.

Let us now consider a more complex example. Consider a right-censored data structure $O = (C, \bar{X}(C))$, where $X = (X(t) : t \in (0, \tau])$ is a time-dependent process representing the full-data structure, C is a right-censoring time, and $\bar{X}(t) = (X(s) : s \leq t)$. Let $R(t) = I(T \leq t)$ be a component of $X(t)$, where T denotes a time to final event of interest, at which time $X()$ is truncated: $X(t) = X(\min(t, T))$. Assume the CAR assumption: $g_0(c \mid X)$ is a function of $(c, \bar{X}(c))$, or equivalently, $\lambda_{g_0}(t \mid X)$ is only a function of $(t, \bar{X}(t))$, where λ_{g_0} is the conditional hazard of censoring, given X. Let Q_0 represent the factor of the density of O under P_0 that only depends on the full-data distribution: $P_0 = Q_0 g_0$, where $Q_0(c, \bar{x}(c)) = P_0(\bar{X}(c) = \bar{x}(c))$.

Consider a particular pathwise differentiable parameter, such as a survival function $\psi_0 = P_0(T > t_0)$ at time t_0. Chapter 3 in van der Laan and Robins (2003) teaches us that the efficient influence curve $D^*(Q_0, g_0)$ can be represented as $D_{IPCW}(g_0, \psi_0) - D_{CAR}(Q_0, g_0)$, where $D_{IPCW}(g_0, \psi_0) = I(T > t_0) I(C > T)/\bar{G}_0(T \mid X) - \psi_0$, $\bar{G}_0(t \mid X) = P_0(C > t \mid X)$, $D_{CAR}(Q_0, g_0) = \int H_{Q_0,g_0}(u) dM_{g_0}(u)$ for a specified function $H_{Q_0,g_0}(u, \bar{X}(u)) = E_0(D_{IPCW}(g_0, \psi_0) \mid \bar{X}(u), C > u)$, and $dM_{g_0}(u) = I(C = u) - I(C \geq u)\lambda_{g_0}(u \mid X)$. Note that $D^*(Q_0, g_0) = D(Q_0, g_0) - \psi_0$ so that we can define the loss function as $L_{g_0}(Q) = D^2(Q, g_0)$. By our general result, we have that the loss-based dissimilarity for Q is given by

$$P_0 D^2(Q, g_0) - P_0 D^2(Q_0, g_0) = P_0 \{D_{CAR}(Q, g_0) - D_{CAR}(Q_0, g_0)\}^2$$
$$= E_0 \int \{H_{Q-Q_0,g_0}(u, \bar{X}(u))\}^2 \lambda_{g_0}(1 - \lambda_{g_0})(du \mid X).$$

Here we essentially used that $T_{CAR}(P_0)$ allows an orthogonal decomposition in $L_0^2(P_0)$ according to the factorization of $g_0(C \mid X) = \prod_t g_0(A(t) \mid \bar{A}(t-), X)$ as a product of conditional distributions of Bernoulli random variables $A(t) = I(C = t)$,

given $(X, \bar{A}(t-) = (I(C = s), s < t))$, and thereby that the variance (i.e., the square of the norm) of an element $D_{CAR}(Q - Q_0, g_0)$ in $T_{CAR}(P_0)$ is a sum of variances. This formula also applies to continuous C through the well-known results for martingales of counting processes (Andersen et al. 1993). Thus, the loss-based dissimilarity is an L^2-norm of $(H_{Q,g_0} - H_{Q_0,g_0})$, where H_{Q_0,g_0} is the principle element that makes up the efficient influence curve. This shows that the super learner $\hat{Q}^*_{\alpha_n}$ will select an estimator that is the best for the purpose of estimating H_{Q_0,g_0}, and thereby the efficient influence curve $D^*(Q_0, g_0)$.

The above result for the loss-based dissimilarity for the loss function L_{g_0} generalizes immediately to causal inference data structures (A, L_A), with A a time-dependent process representing censoring and treatment actions (i.e., the intervention nodes), L_A a time-dependent process including time-dependent covariates and outcomes, and L_a the counterfactual corresponding with a multiple-time-point intervention that sets A equal to the treatment profile a.

A.23 Example: EEM and TMLE

Let us revisit the missing outcome example with $O = (W, \Delta, \Delta Y) \sim P_0 \in \mathcal{M}$, $\mathcal{M}$ the nonparametric model, and $\psi_0 = E_0 Y$. Let $\Pi_0(W) = P_0(\Delta = 1 \mid W) = g_0(1 \mid W)$. The efficient influence curve is given by

$$D^*(Q_0, \Pi_0)(O) = \Delta/\Pi_0(W)(Y - \bar{Q}_0(W)) + \bar{Q}_0(W) - \Psi(Q_0),$$

where $\bar{Q}_0(W) = E_0(Y \mid W, \Delta = 1)$, $Q_0 = (Q_{W,0}, \bar{Q}_0 = E_0(Y \mid W, \Delta = 1))$. Note $D^*(Q_0, \Pi_0) = D(\bar{Q}_0, \Pi_0) - \psi_0$, . Given a parametric family $\mathcal{Q}^w$ for $\bar{Q}_0$, as shown in Rubin and van der Laan (2008) and above, minimizing $E_0 D^2(\bar{Q}, g_0)$ over $\bar{Q} \in \mathcal{Q}^w$ corresponds with

$$\arg\min_{\bar{Q} \in \mathcal{Q}^w} E_0 \frac{\Delta(1 - \Pi_0)}{\Pi_0^2}(Y - \bar{Q}(W))^2.$$

TMLE is a substitution estimator and thus has advantages over other asymptotically efficient estimators. We now want to combine TMLE with EEM, so that we can also claim that TMLE is asymptotically linear with influence curve that is optimal among a given class of influence curves $D^*(\bar{Q}, \Pi_0, \psi_0)$ with $\bar{Q} \in \mathcal{Q}^w$. We consider EEM for the linear and logistic TMLE with respect to a parametric family $\{\bar{Q}_\alpha : \alpha\}$. The linear TMLE provides a closed-form algebraic demonstration, but does not respect known bounds, so that the preferred TMLE is the logistic TMLE, which follows. Let $Y \in [0, 1]$. Let $H_{\Pi_0} = H_0 = \frac{\Delta}{\Pi_0}$.

TMLE, squared error loss, linear fluctuation. Consider the linear fluctuation $\bar{Q}_\alpha(\epsilon) = \bar{Q}_\alpha + \epsilon H_{\Pi_0}$. Define

$$\epsilon_n(\alpha) = \arg\min_{\epsilon} P_n L_2(\bar{Q}_\alpha(\epsilon)),$$

where $L_2(\bar{Q})(O) = \Delta(Y - \bar{Q}(W))^2$. Note $\epsilon_n(\alpha)$ is the univariate linear regression coefficient (no intercept) of $(Y - \bar{Q}_\alpha)$ on H_{Π_0}. Thus

$$\epsilon_n(\alpha) = \frac{E_{P_n}\Delta(Y - \bar{Q}_\alpha)H_{\Pi_0}}{E_{P_n}\Delta H_{\Pi_0}^2}.$$

The candidate TMLEs are defined as $\bar{Q}^*_{\alpha,n} = \bar{Q}_\alpha + \epsilon_n(\alpha)H_{\Pi_0}$. We wish to determine α so that

$$\alpha \to E_0 D^2(\bar{Q}^*_{\alpha,n}, \Pi_0)$$

is minimized. Directly minimizing the empirical counterpart

$$\alpha \to P_n D(Q_\alpha, g_0)^2 = E_{P_n}\left\{Y\frac{\Delta}{\Pi_0} - \bar{Q}^*_{\alpha,n}(W)\left(\frac{\Delta}{\Pi_0} - 1\right)\right\}^2$$

corresponds with an unweighted least squares regression of an inverse-weighted outcome on an inverse-weighted corrected covariate. By the above result, this choice can also be estimated as

$$\alpha_n = \arg\min_\alpha E_{P_n}\frac{\Delta(1 - \Pi_0)}{\Pi_0^2}(Y - \bar{Q}^*_{\alpha,n}(W))^2.$$

The resulting TMLE is given by $\bar{Q}^*_{\alpha_n,n} = \bar{Q}_{\alpha_n} + \epsilon_n(\alpha_n)H_{\Pi_0}$. Note,

$$\begin{aligned}
\alpha_n &= \arg\min_\alpha E_{P_n}\frac{\Delta(1 - \Pi_0)}{\Pi_0^2}(Y - \bar{Q}_\alpha - \epsilon_n(\alpha)H_0)^2 \\
&= \arg\min_\alpha E_{P_n}\frac{\Delta(1 - \Pi_0)}{\Pi_0^2}\left(Y - \bar{Q}_\alpha - \frac{E_{P_n}\Delta(Y - \bar{Q}_\alpha)H_0}{E_{P_n}\Delta H_0^2}H_0\right)^2 \\
&= \arg\min_\alpha E_{P_n}\frac{\Delta(1 - \Pi_0)}{\Pi_0^2}\left(Y - \frac{E_{P_n}\Delta Y H_0}{E_{P_n}\Delta H_0^2}H_0 - \bar{Q}_\alpha + \frac{E_{P_n}\Delta\bar{Q}_\alpha H_0}{E_{P_n}\Delta H_0^2}H_0\right)^2.
\end{aligned}$$

If $\bar{Q}_\alpha = \alpha W$ is linear, then it follows that α_n is a weighted-linear-least-squares estimator regressing $Y - \frac{E_{P_n}\Delta Y H_0}{E_{P_n}\Delta H_0^2}H_0$ on the covariate

$$W' \equiv W - \frac{E_{P_n}\Delta W H_0}{E_{P_n}\Delta H_0^2}H_0$$

according to a linear model $\alpha W'$.

TMLE, quasi-log-likelihood loss, logistic fluctuation. The above TMLE behaves poorly under violations of the positivity assumption, since the linear fluctuation does not respect bounds, for example, $Y \in [a, b]$ for some values $[a, b]$. Therefore, we proposed an alternative TMLE based on the binary-log-likelihood loss function and logistic fluctuation. Let $\{\bar{Q}_\alpha : \alpha\}$ be a logistic regression family. Consider $\text{logit}\bar{Q}_\alpha(\epsilon) = \text{logit}\bar{Q}_\alpha + \epsilon H_0$. Define

$$\epsilon_n(\alpha) = \arg\min_\epsilon P_n L(\bar{Q}_\alpha(\epsilon)),$$

where $L(\bar{Q})(O) = \Delta\left\{Y \log \bar{Q}_\alpha + (1-Y)\log(1-\bar{Q}_\alpha)\right\}$. Note $\epsilon_n(\alpha)$ is the univariate linear logistic regression coefficient Y on H_0 using $\text{logit}\bar{Q}_\alpha$ as intercept. We have that $\epsilon_n(\alpha)$ solves

$$0 = \textstyle\sum_i \Delta_i H_0(W_i)(Y_i - \bar{Q}_\alpha(\epsilon)(W_i)).$$

Even though $\epsilon_n(\alpha)$ is not a closed form function of α, this equation allows us to determine closed-form expressions for first-order derivatives $\frac{d}{d\alpha}\epsilon_n(\alpha)$. If $\epsilon(\alpha)$ is defined as $U(\alpha, \epsilon(\alpha)) = 0$ for an equation U, then

$$\frac{d}{d\alpha}\epsilon(\alpha) = -\left\{\frac{d}{d\epsilon}U(\alpha,\epsilon)\Big|_{\epsilon=\epsilon(\alpha)}\right\}^{-1}\frac{d}{d\alpha}U(\alpha,\epsilon).$$

The candidate TMLEs are defined as $\text{logit}\bar{Q}^*_{\alpha,n} = \text{logit}\bar{Q}_\alpha + \epsilon_n(\alpha)H_0$. We want to determine the minimizer of

$$\alpha \to P_0 D^2(\bar{Q}^*_{\alpha,n}, \Pi_0).$$

This choice can be estimated as

$$\alpha_n = \arg\min_\alpha E_{P_n}\frac{\Delta(1-\Pi_0)}{\Pi_0^2}(Y - \bar{Q}^*_{\alpha,n}(W))^2.$$

The desired TMLE is given by $\bar{Q}^*_{\alpha_n,n} = \bar{Q}_{\alpha_n}(\epsilon_n(\alpha_n))$. Solving for α_n corresponds with a nonlinear least squares problem. Fast algorithms for solving for this α_n will require (1) fast evaluation of $\epsilon_n(\alpha)$ and (2) closed-form expression for the derivatives in α. Since we have closed-form derivatives of $\alpha \to \epsilon_n(\alpha)$ this can be carried out with available software.

TMLE logistic with linear regression family. Suppose that we want to optimize efficiency over a linear regression model $\bar{Q}_\alpha = \alpha W$ instead of the logistic linear regression model above. We could use the optimal α_n as defined for the linear TMLE. This now defines an initial linear-least-squares estimator $\bar{Q}_{\alpha_n}$. We can truncate this fit between $(0, 1)$ and compute the logistic TMLE update.