University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2012 *Paper* 300

Causal Inference for Networks

Mark J. van der Laan*

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

Causal Inference for Networks

Mark J. van der Laan

Abstract

Suppose that we observe a population of causally connected units according to a network. On each unit we observe a set of potentially connected units that contains the true connections, and a longitudinal data structure, which includes time-dependent exposure or treatment, time-dependent covariates, a final outcome of interest. The target quantity of interest is defined as the mean outcome for this group of units if the exposures of the units would be probabilistically assigned according to a known specified mechanism, where the latter is called a stochastic intervention. Causal effects of interest are defined as contrasts of the mean of the unit specific outcomes under different stochastic interventions one wishes to evaluate. By varying the network structure, this covers a large range of estimation problems ranging from independent units, independent clusters of units, anda single cluster of units in which each unit has a limited number of connections to other units. We present a few motivating classes of examples, propose a structural causal model, define the desired causal quantities, address the identification of these quantities from the observed data, and define maximum likelihood based estimators based on cross-validation.

Such smoothed/regularized maximum likelihood estimators are not targeted and will thereby be overly bias w.r.t. the target parameter, and, as a consequence, generally not result in asymptotically normally distributed estimators of the statistical target parameter. Therefore, we formulated targeted maximum likelihood estimators of this estimand, and showed that the robustness of the efficient influence curve implies that the bias of the TMLE will be a second order term involving squared differences of two nuisance parameters. In order to deal with the curse of dimensionality, we present super-learning based on cross-validation, and we develop targeted maximum likelihood estimators, which are less biased than maximum likelihood estimators due to a targeted bias reduction step. Due to the causal dependencies between units, the data set may correspond with the realization of

a single experiment, so that establishing a (e.g., normal) limit distribution for the estimators, and corresponding statistical inference, is a challenging topic. In order to establish a formal theorem, we focus on the point-treatment longitudinal data structure, thereby also putting down a foundation for its generalization to the general longitudinal data structure, which we reserve for future research.We conclude with a discussion.

1 Introduction and motivation

Most of the literature on causal inference has focussed on assessing the causal effect of a single or multiple time-point intervention on some outcome based on observing n longitudinal data structures on n independent units that are not causally connected. For literature reviews we refer to a number of books on this topic: Rubin (2006), Pearl (2009), van der Laan and Robins (2003), Tsiatis (2006), Hernán and Robins (2012), van der Laan and Rose (2012).

Such a causal effect is defined as an expectation of the direct effect of the intervention assigned to the unit on the unit's outcome, since indirect causal effects on the unit's outcome through other units are assumed non-existent. As a consequence, causal models only have to be concerned about the modeling of causal relations between the components of the unit-specific data structure. Statistical inference is based on the assumption that the n data structures can be viewed as n independent realizations of a random variable, so that central limit theorems for sums of independent random variables can be employed. The latter requires that the sample size n is large enough so that statistical inference based on the normal limit distributions is indeed appropriate.

In many applications one may define the unit as a group of causally connected individuals, often called a community or cluster. It is then assumed that the communities are not causally connected, and that the community specific data structures can be represented as n independent random variables. One can then define a community specific outcome, and assess the causal effect of the community level intervention/exposure on this community specific outcome with methods from the causal inference literature. Such causal effects incorporate indirect as well as direct effects of the community level intervention, where the indirect effects of the community level exposure on an individual in a community occur through other individuals in that same community. We refer to Halloran and Struchiner (1995); Hudgens and Halloran (2008); VanderWeele et al. (2012); Tchetgen Tchetgen and VanderWeele (2012) for defining different types of causal effects in the presence of causal interference between units. We also refer to Donner and Klar (2000), Hayes and Moulton (2009), Campbell et al. (2007b) for reviews on cluster randomized trials and cluster level observational studies.

In many such community randomized trials or observational studies the number of communities is very small (e.g., around 10 or so), so that the number of independent units itself is not large enough for statistical inference based on limit distributions. In the extreme, but not uncommon, case, one may observe a single community of causally connected individuals. Can one now still statistically evaluate a causal effect of an intervention assigned at the community level on a community level outcome, such as the average of individual outcomes? Clearly, causal models incorporating all units are needed in order to define the desired causal quantity, and identifiability of these causal quantities under (minimal) assumptions need to be established without relying on asymptotics in a number of *independent* units.

An important ingredient of our modeling approach carried out in this article is the incorporation of network information that describes for each unit (in a finite population of N units) a set of other units this unit is potentially connected to. The more precise network information will be available, the larger the effective sample size will be for targeting the desired quantity. We will also assume sequential conditional independence of the units at time t, conditional on the past of all units at time t. That is, conditional on the most recent past on all units, including the recent network information, the data on the units at the next time point are independent across units. Even though this network information allows the units to depend on each other in complex ways, we will demonstrate that the likelihood of the data on all N units allows statistical inference driven by the number of units instead of driven by the number of communities (e.g., 1).

We will apply the roadmap for targeted learning of a causal effect (e.g., van der Laan and Rose (2012), Pearl (2009), Petersen and van der Laan (2012)), which starts out with defining a structural causal model, defining the causal quantity of interest, and defining the observed data and its link to the data generated by the causal system. Subsequently, we then establish the identifiability of the causal quantity from the data distribution under transparent additional (often non-testable) assumptions. This identifiability result allows us to define a statistical model that contains the true probability distribution of the data, and an estimand (i.e. a target parameter mapping applied to true data distribution) that reduces to this causal quantity if the required causal assumptions hold. This statistical model, and the target parameter mapping that maps data distributions in this statistical model into the parameter values, defines the estimation problem. Finally, we have to develop targeted estimators of the target parameter and develop the theory for statistical inference. The statistical model needs to contain the true data distribution, so that the statistical estimand can be interpreted as a pure statistical target parameter, while under the stated additional causal conditions that were needed to identify the causal effect, it can be interpreted as the causal quantity of interest. To understand the deviation between the estimand and the causal quantity under a variety of violations of these causal assumptions, one may carry out a sensitivity type analysis.

Since the statistical model does not assume that the data generating experiment involves the repetition of independent experiments, the development of targeted estimators and inference represents novel and new challenges in estimation and inference that, to the best of our knowledge, have not been earlier attacked by the current causal inference literature. Targeted minimum loss-based estimation was developed for estimation in semi-parametric models for i.i.d. data (van der Laan and Rubin, 2006; van der Laan, 2008; van der Laan and Rose, 2012), and extended to a particular form of dependent treatment/censoring allocation as present in group sequential adaptive designs (van der Laan, 2008; Chambaz and van der Laan, 2011a,b; van der Laan and Petersen, 2012). In this article we need to generalize targeted minimum loss based estimation to the complex semiparametric statistical model presented in this article, and we also need to develop corresponding statistical inference.

Our models generalize the models in the causal inference literature for independent units, and thus avoid assumptions that would not be needed when applying the methods to observing data on units that are causally and statistically independent. In particular, our models include causal inference for community based interventions based on observing a number of independent communities/clusters. In addition, our models also incorporate group sequential adaptive designs in which treatment allocation to an individual can be based on what has been observed on previously recruited individuals in the trial (Hu and Rosenberger, 2006; van der Laan, 2008; Chambaz and van der Laan, 2011a,b; van der Laan and Petersen, 2012). Our models also allow that the outcome of an individual is a function of the treatments other individuals received The latter is referred to as interference in the causal inference literature. Thus the causal models proposed in this article do not only generalize the existing causal models for independent units, but they also generalize causal models that incorporate previously studied causal dependencies between units. Finally, we note that our models and corresponding methodology can be used to establish a methodology for assessing causal effects of interventions on the network on the average of the unit specific outcomes. For example, one might want to know how the community level outcome changes if we change the network structure of the community through some intervention, such as increasing the connectivity between certain units in the community?

1.1 Organization of article

The organization of this article is as follows.
Section 2: We formulate a counterfactual causal model that can be viewed as an analogue of the structural causal model actually used in this article. This

section provides a perspective of the contribution of this article in the context of the causal inference literature that relies on the Neyman-Rubin model, demonstrating that in essence it corresponds with simultaneously allowing for causal interference between the units and that interventions assigned to a unit are informed by other units in the population.

Section 3: We present our structural causal model, stochastic interventions and corresponding counterfactual outcomes, causal quantity defined in terms of the mean of the intervention specific counterfactuals, identifiability of causal quantity from data distribution P_0 of data $O = (O_1, \ldots, O_N)$ on the N units, statistical model $\mathcal{M}$ for the probability distribution of O, statistical target parameter mapping $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ that defines the estimand $\Psi(P_0)$, where the latter reduces to the causal quantity under the additional assumptions that were needed to establish the identifiability. The statistical estimation problem is now defined by the data $O \sim P_0 \in \mathcal{M}$, the statistical model $\mathcal{M}$ and target parameter $\Psi : \mathcal{M} \rightarrow \mathbb{R}$. The likelihood of O factorizes as $P = Qg$ and $\Psi(P)$ only depends on P through the Q-factor. Therefore, we also use the notation $\Psi(Q)$ to denote this target parameter $\Psi(P)$.

Section 4: We present some general classes of examples covered by our causal models and causal quantities.

Section 5: We discuss maximum likelihood estimation (MLE), unified loss-based cross-validation (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006), and likelihood based super learning (van der Laan et al., 2007; Polley et al., 2012). Such smoothed/regularized maximum likelihood estimators are not targeted and will thereby be overly bias w.r.t. the target parameter, and, as a consequence, generally not result in asymptotically normally distributed estimators of the statistical target parameter. Thus there is a need for targeted learning (targeting the fit towards ψ_0) instead of MLE in order to deal with the curse of dimensionality.

Section 6: We derive the efficient influence curve, also called the canonical gradient of the pathwise derivative of the statistical target parameter (Bickel et al. (1997); van der Vaart (1998)). We also establish that the expectation of the efficient influence curve $D^*(Q, g)$ under misspecified parameters (Q, g) of the data distribution $P = Q * g$ can be represented as $\Psi(Q_0) - \Psi(Q)$ plus a sum of two second order terms, one involving square differences between Q and Q_0, and another a product of differences of Q and Q_0 and a specified $h(Q, g)$ and $h(Q_0, g_0)$. This result provides a fundamental ingredient in establishing a first order expansion of the targeted maximum likelihood estimator (TMLE) under conditions that make these second order terms negligible relative to the first order term, while a separate analysis of the first order term (which is a sum of dependent random variables) establishes the asymptotic normality of the TMLE.

Section 7: Our heuristic arguments demonstrate that the log-likelihood of O will satisfy a local asymptotic normality condition (Bickel et al. (1997); van der Vaart (1998)) so that efficiency theory can be applied to path wise differentiable target parameters of the data distribution. As demonstrated in van der Vaart (1998), under local asymptotic normality the normal limit distribution of the MLE (ignoring all regularity conditions that would be needed to establish the asymptotic normality of the MLE) is optimal in the sense of the convolution theorem. In this section 7 we demonstrate that the variance of the efficient influence curve (i.e., the canonical gradient of the pathwise derivative of the target parameter) corresponds with the asymptotic variance of a maximum likelihood estimator of the target parameter. From this we learn that our goal should be to construct estimators that are asymptotically normally distributed with variance equal to the standardized variance of the efficient influence curve (and thus asymptotically equivalent with a MLE), while appropriately dealing with the curse of dimensionality through super learning and targeted maximum likelihood estimation (TMLE).

Section 8: We present a general TMLE based on least-favorable submodels implied by the efficient influence curve, analogue to the TMLE for i.i.d. data (van der Laan and Rubin (2006)).

Section 9: We present the TMLE for the causal effect of a single time point intervention on an outcome, controlling for the baseline covariates across the units, an important special case. This TMLE generalizes the TMLE of the causal effect of a single time point intervention under causal and statistical independence of the units (Gruber and van der Laan (2010); Rosenblum and van der Laan (2010); van der Laan and Rose (2012). It is shown that the efficient influence curve satisfies a double robustness property, which implies the double robustness of the TMLE. Interestingly, it is demonstrated that defining the estimator as the solution of the efficient influence curve estimating equation fails as a method, by not being able to utilize the double robustness of the efficient influence curve, while the TMLE still inherits the double robustness of the efficient influence curve. This provides another demonstration of the importance of TMLE relative to estimating equation methodology Robins and Rotnitzky (1992); van der Laan and Robins (2003) that defines the estimator as a solution of an estimating equation.

Section 10: We present a theorem establishing asymptotic normality of this TMLE for the causal effect of a single time point intervention, and discuss statistical inference based on its normal limit distribution. The theorem relies on modern advances in weak convergence of processes as presented in van der Vaart and Wellner (1996); van der Vaart (1998). The proof of the theorem

is deferred to the Appendix. The generalization of the formal asymptotics results for this TMLE to the TMLE for general longitudinal data structures is also discussed in the final subsection of the Appendix.
Section 11: We extend the latter TMLE and inference under a larger model that weakens the sequential conditional independence assumption. **Section 12:** We discuss the extension of our results to the case that our observed data corresponds with a random sample of individuals from the complete network of individuals.
Section 13: We conclude with a summary and some concluding remarks.

2 Possible formulation of estimation problem in terms of counterfactuals

The estimation problem defined in the next section in terms of a semi-parametric structural equation model corresponds with the following counterfactual missing data problem formulation also called the Neyman-Rubin causal model (Neyman (1990); Rubin (1974, 2006); Holland (1986); Robins (1987a,b)).

Let $X_i^F = (L_{i,a} : a \in \mathcal{A})$ be the full-data structure consisting of all static regimen specific counterfactuals for unit i, where $a = (a_1, \ldots, a_N)$ represents the static regimens for all N units, $L_{i,a} = (L_{i,a}(0), L_{i,a}(1), \ldots, L_{i,a}(\tau+1))$ is a time-dependent process up till time $\tau + 1$, and $L_{i,a}(t)$ only depends on a through $(\bar{a}_j(t-1) = (a_j(0), \ldots, a_j(t-1)) : j = 1, \ldots, N)$. We assume X_i^F, $i = 1, \ldots, N$, are independent and identically distributed, or, most importantly, independently distributed.

Let P_0^F be the probability distribution of X^F and let $\mathcal{M}^F$ be the full-data model, i.e., the collection of possible distributions of X^F. This full data model will thus incorporate additional assumptions such as that the counterfactuals of unit i only depend on the regimens of a subset of the N individuals. We observe the missing data structure $O = (O_i : i = 1, \ldots, N)$, $O_i = (A, L_i = L_{i,A})$ on the full data X_i^F, $i = 1, \ldots, N$. We view $O = (O_1, \ldots, O_N)$ as a missing data structure on the full-data $X^F = (X_1^F, \ldots, X_N^F)$. We assume that the conditional density g_0 of $A = (A_1, \ldots, A_N)$, given X^F, satisfies $g_0(A \mid X^F) = \prod_{t=0}^{\tau} \prod_{i=1}^{N} g_{0,t,i}(A_i(t) \mid c_{t,i}^A)$, where $c_{t,i}^A$ is a function of $(\bar{A}_j(t-1), \bar{L}_j(t) : j = 1, \ldots, N)$, so that the missingness mechanism satisfies coarsening at random. Note that $g_0(A \mid X^F) = h_0(O)$ is a function of O so that this assumption indeed implies the coarsening at random assumption. Due to this coarsening at random assumption, the likelihood of O factorizes:

$$P_0(A, L_i : i = 1, \ldots, N) = \prod_{i=1}^{N} P_{P_0^F}(L_{i,a} = l_i)\Bigg|_{a=A, l_i=L_i} g_0(A \mid X^F).$$

Note that the full-data distribution factor equals the likelihood of $(L_{i,a} : i = 1, \ldots, N)$ (i.e., vector of independent counterfactuals) at set regimen $a = (a_1, \ldots, a_N)$ at value $A = (A_1, \ldots, A_N)$, and is thus identified by the full-data distribution P_0^F.

Our target parameter is a parameter $\Psi^F : \mathcal{M}^F \rightarrow \mathbb{R}^d$ defined on the full-data model. The factorization of the likelihood of O due to CAR establishes the identifiability of $\psi_0^F = \Psi^F(P_0^F)$ as a parameter of the distribution P_0 of O, under the assumption that $\Psi^F(P_0^F)$ only depends on P_0^F through the relevant factor of the likelihood. As a consequence, we can now define a statistical target parameter $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ so that $\psi_0^F = \psi_0 = \Psi(P_0)$. We need to construct an estimator of ψ_0 based on this single draw of $O \sim P_0 \in \mathcal{M}$, and we need to establish a limit distribution of this standardized estimator $\sqrt{N}(\psi_N - \psi_0) \Rightarrow_d Z$ for some limit distribution Z.

The fact that the counterfactual outcomes of subject i can be a function of the treatments of other subjects is often referred to as interference in the causal inference literature. In addition, the above formulation allows that treatment allocation for unit i depends on data collected on other units. The above formulation can thus be viewed as the causal inference estimation problem when interference and adaptive treatment allocation is allowed. Our structural equation model defined in next section implies restrictions on the distribution of the counterfactuals, and thus defines a particular full-data model $\mathcal{M}^F$.

3 Formulation of estimation problem.

Suppose we have a population of interconnected units that evolves over time. Let $(X_i(t) : t \in \{0, 1, \ldots, \tau + 1\})$, $i = 1, \ldots, N$, be the time-dependent processes associated with each of N units. For a unit i, we define $A_i(t) \subset X_i(t)$ as an action at time t, which will play the role of intervention node in the structural equation model below. The process on individual i at time t, $X_i(t)$ consists of measurements $L_i(t)$, possibly including an outcome process $Y_i(t)$, the intervention node $A_i(t)$, while $L_i(\tau + 1) = Y_i$ denotes a final outcome. Let $F_i(t) \subset L_i(t)$ denote the network of friends individual i is potentially connected to at time t, $t = 0, \ldots, \tau + 1$.

In our causal model we will assume that each unit's $X_i(t)$ can only be a function of the past $\bar{X}(t-1) \equiv (\bar{X}_j(t-1) : j = 1, \ldots, N)$, of all subjects, where $\bar{X}_j(t-1) = (X_j(s) : s \leq t-1)$: i.e., the units $X_1(t), \ldots, X_N(t)$ are conditionally independent, given $\bar{X}(t-1)$. Below, we make some remarks regarding weakening this assumption.

If we define $L(t) = (L_i(t) : i = 1, \ldots, N)$, and similarly we define $A(t)$, then the ordered single data structure can be represented as

$$(L(0), A(0), \ldots, L(\tau), A(\tau), Y = L(\tau+1)).$$

The latter ordering is the only causally relevant ordering, and the ordering of units within a time-point is user supplied but inconsequential. We define $Pa(A(t)) = (\bar{L}(t), \bar{A}(t-1))$, $Pa(L(t)) = (\bar{L}(t-1), \bar{A}(t-1))$, as the parent nodes of $A(t)$ and $L(t)$, respectively, w.r.t. this ordering. Note that the parent nodes of $A_i(t)$, denoted with $Pa(A_i(t))$, equal $Pa(A(t))$, and, the parent nodes of $L_i(t)$, denoted with $Pa(L_i(t))$, equal $Pa(L(t))$, $t = 0, \ldots, \tau{+}1$, $i = 1, \ldots, N$.

This ordered sequence for $X = (X_1, \ldots, X_N)$ and the specification of the parent-nodes implies a corresponding set of structural equations: first generate $U_N = (U_i : i = 1, \ldots, N)$ where

$$U_i = (U_{L_i(0)}, U_{A_i(0)}, \ldots, U_{L_i(\tau)}, U_{A_i(\tau)}, U_{Y_i}), \quad i = 1, \ldots, N,$$

and then generate X deterministically as follows:

$$\begin{aligned}
&L_i(t) = f_{L_i(t)}(Pa(L_i(t)), U_{L_i(t)}) \\
&A_i(t) = f_{A_i(t)}(Pa(A_i(t)), U_{A_i(t)}) \\
&i = 1, \ldots, N, t = 0, \ldots, \tau \\
&Y_i = f_{Y_i}(Pa(Y_i(\tau+1)), U_{Y_i(\tau+1)}) \\
&i = 1, \ldots, N.
\end{aligned}$$

Since $Pa(L_i(t)) = (\bar{A}(t-1), \bar{L}(t-1))$ and $Pa(A_i(t)) = (\bar{A}(t-1), \bar{L}(t))$, an alternative succinct way to represent this structural equation model is:

$$\begin{aligned}
&L(t) = f_{L(t)}(Pa(L(t)), U_{L(t)}) \\
&A(t) = f_{A(t)}(Pa(A(t)), U_{A(t)}) \\
&t = 0, \ldots, \tau \\
&Y = L(\tau+1) = f_Y(Pa(Y), U_Y).
\end{aligned}$$

Counterfactuals and stochastic interventions: This structural equation model for

$$(L(0), A(0), \ldots, L(\tau), A(\tau), Y = L(\tau+1)),$$

allows us to define counterfactuals $Y_{\mathbf{d}}(\tau+1)$ corresponding with an dynamic intervention d on A (Robins (1987b,b, 1997, 1999); Gill and Robins (2001);

Yu and van der Laan (2003). For example, one could define $A_i(t)$ at time t as a particular deterministic function $d_{i,t}$ of the parents $Pa(A_i(t))$ of subject $i = 1, \ldots, N$. Such an intervention corresponds with replacing the equations for $A(t)$ by this deterministic equation $d_t(Pa(A(t))$, $t = 0, \ldots, \tau$. More generally, we can replace the equations for $A(t)$ that describe a degenerate distribution for drawing $A(t)$, given $U = u$, and $Pa(A(t))$, by a user supplied conditional distribution of an $A^*(t)$, given $Pa(A^*(t))$. Such a conditional distribution defines a so called stochastic intervention Dawid and Didelez (2010); Didelez et al. (2006); Diaz and van der Laan (2012).

Let $g^* = (g_t^* : t = 0, \ldots, \tau)$ denote our selection of a stochastic intervention identified by a set of conditional distributions of $A^*(t)$, given $Pa(A^*(t))$, $t = 0, \ldots, \tau$. For convenience, we represent the stochastic intervention with equations $A^*(t) = f_{A^*(t)}(Pa(A^*(t)), U_{A^*(t)})$ in terms of random errors $U_{A^*(t)}$. This implies the following modified system of structural equations:

$$\begin{aligned}
&L(t) = f_{L(t)}(Pa(L(t)), U_{L(t)})\\
&A^*(t) = f_{A^*(t)}(Pa(A^*(t)), U_{A^*(t)})\\
&t = 0, \ldots, \tau\\
&Y = L(\tau + 1) = f_Y(Pa(Y), U_Y).
\end{aligned}$$

Let Y_{i,g^*}, or short-hand $Y_{i,*}$, denote the corresponding counterfactual outcome for unit i. A causal effect at the unit level could now be defined as a contrast such as $Y_{i,g_1^*} - Y_{i,g_2^*}$ for two interventions g_1^* and g_2^*. Note that, for a given g^*, $Y_{i,g^*} = Y_{i,g*}(U^*)$ is a deterministic function of the error-term $U^* = (U, U_A^*)$ that are inputted in the structural equations. In particular, for deterministic interventions g^* Y_{i,g^*} is a deterministic function of the error terms U.

Post-intervention distribution, and sequential randomization assumption: If we assume the sequential randomization assumption on U, i.e., $A(t)$ is independent of L_{g^*}, given $Pa(A(t))$, and U^* is externally generated so that $U^* \perp U$, then the probability distribution P_{g^*} of (A^*, L_{g^*}) is given by the so called G-computation formula (Robins (1987a); Gill and Robins (2001); Yu and van der Laan (2003); Didelez et al. (2006); Zheng and van der Laan (2012))

$$P_{g^*}(A^*, L) \;=\; \prod_{t=0}^{\tau+1} \prod_{i=1}^{N} P_{L_i(t)}(L_i(t) \mid Pa^*(L_i(t))) g_t^*(A_i^*(t) \mid Pa(A_i^*(t))),$$

where $P_{L_i(t)}$ is the conditional distribution of $L_i(t)$, given $Pa(L_i(t))$, and $Pa^*(L_i(t)) = (\bar{L}(t-1), \bar{A}^*(t-1))$. Thus, the post-intervention distribution P_{g^*} is identified from the distribution of $X = (A, L)$ generated by the structural equation model. We will assume this sequential randomization assumption. The distribution of Y_{i,g^*} corresponds now with a marginal distribution of $P_{L_{g^*}}$.

Average causal effect (ACE): One might now define an average causal effect as the following target parameter of this distribution of P_{g^*}:

$$\psi^F = E_{P_{g_1^*}} \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_{i,g_1^*} \right\} - E_{P_{g_2^*}} \left\{ \frac{1}{N} \sum_{i=1}^{N} Y_{i,g_2^*} \right\}.$$

Let $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$. Since the distribution P_{g^*} is indexed by N, the parameter ψ^F can depend on N.

Iterative conditional expectation representation of ACE: The parameter $E\bar{Y}_{g^*}$ can be represented as an iterative conditional expectation w.r.t. the probability distribution P_{g^*} of (A^*, L_{g^*}) (Bang and Robins (2005); van der Laan and Gruber (2012)):

$$\begin{array}{l}
\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i(\tau+1) \\
\bar{Q}^{g^*}_{\tau+1,1} = E(\bar{Y} \mid \bar{L}(\tau), \bar{A}(\tau)) \\
\bar{Q}^{g^*}_{\tau+1} = E_{g^*_\tau}(\bar{Q}^{g^*}_{\tau+1,1} \mid \bar{L}(\tau), \bar{A}(\tau-1)) \\
\bar{Q}^{g^*}_{\tau,1} = E(\bar{Q}^{g^*}_{\tau+1} \mid \bar{L}(\tau-1), \bar{A}(\tau-1)) \\
\bar{Q}^{g^*}_{\tau} = E_{g^*_{\tau-1}}(\bar{Q}^{g^*}_{\tau,1} \mid \bar{L}(\tau-1), \bar{A}(\tau-2)) \\
\text{Iterate} \\
\bar{Q}^{g^*}_{1,1} = E(\bar{Q}^{g^*}_{2} \mid \bar{L}(0), \bar{A}(0)) \\
\bar{Q}^{g^*}_{1} = E_{g^*_0}(\bar{Q}^{g^*}_{1,1} \mid \bar{L}(0)) \\
\bar{Q}^{g^*}_{0} = E_{L(0)} \bar{Q}^{g^*}_{1}(L(0)),
\end{array}$$

where $EY_{g^*} = \bar{Q}^{g^*}_0$. Thus, this mapping involves iteratively integrating w.r.t. the observed data distribution of $L(t)$, given its parents, and the conditional distribution g^*_t of $A^*(t)$, given $Pa(A^*(t))$, respectively, starting at $t = \tau + 1$, till $t = 0$.

Dimension reduction and exchangeability across units: We will also assume that for each node $A_i(t)$, $L_i(t)$, we can define functions, $Pa(A_i(t)) \rightarrow c^A_{t,i}(Pa(A_i(t)))$, $Pa(L_i(t)) \rightarrow c^L_{t,i}(Pa(L_i(t)))$, that map into $\mathbb{R}^d$ for a d that does not depend on N, and corresponding common (in i) functions $f_{L(t)}, f_{A(t)}, f_Y$, so that

$$\begin{array}{l} L_i(t) = f_{L(t)}(c^L_{t,i}(Pa(L_i(t))), U_{L_i(t)}) \\ A_i(t) = f_{A(t)}(c^A_{t,i}(Pa(A_i(t))), U_{A_i(t)}) \\ i = 1, \ldots, N, t = 0, \ldots, \tau \\ Y_i = f_Y(c^Y_{\tau+1,i}(Pa(Y_i)), U_{Y_i}) \\ i = 1, \ldots, N. \end{array}$$

Examples of such dimension reductions are $c^L_{t,i}(Pa(L_i(t))) = (\bar{L}_j(t-1), \bar{A}_j(t-1) : j \in F_i(t-1))$, i.e., the observed past of unit i itself and the observed past of its current friends $F_i(t-1)$, and, similarly, we can define $c^A_{t,i}(Pa(A_i(t))) = (\bar{L}_j(t), \bar{A}_j(t-1) : j \in F_i(t-1))$. By augmenting these reductions to data on maximally K friends, filling up the empty cells for units with fewer than K friends with a missing value, these dimension reductions have a fixed dimension (say) d, and include the information on the number of friends. This structural equation model assumes that, across all units i, the data on unit i at the next time point t is a common function of its own past and past of its friends. We will also assume that $U_{L_i(0)}$, $i = 1, \ldots, N$, are independent, so that $L_i(0)$, $i = 1, \ldots, N$, are independent, and we will assume that all the other components of U_i are not only independent across i, but are also identically distributed.

Thus all the dependence between units is not due to the dependence of the errors, but only due to the interdependence between units as described by the system that allows that the unit's data at time t is a function of the data of its friends. In addition, each unit i satisfies the same laws as functions of the i-specific past (representing the unit's past and past of its friends) for generating their future data points, with the exception of $L_i(0)$, which we only require to be independently distributed across the N units.

Identifiability: G-computation formula for stochastic intervention. For notational convenience, let $c^L_{t,i} = c^L_t(Pa(L_i(t)))$, and let $c^{L,*}_{t,i}$ be defined accordingly with A replaced by A^*. Due to the exchangeability and dimension reduction assumptions, the probability distribution P_{g^*} of $L_{g^*} = (L_{i,g^*} : i = 1, \ldots, N)$ now simplifies:

$$\begin{aligned} P_{g^*}(L, A^*) &= \prod_{i=1}^N P_{L_i(0)}(L_i(0)) \prod_{i=1}^N \prod_{t=0}^{\tau+1} P_{L(t)}(L_i(t) \mid c^{L,*}_{t,i}) g^*_t(A^*_i(t) \mid Pa(A^*_i(t))) \tag{1} \\ &\equiv P^{g^*}(L, A^*), \tag{2} \end{aligned}$$

where $P_{L(t)}$ are the above defined conditional distributions of $L_i(t)$, given $Pa(L_i(t))$, and these i-specific conditional densities are constant in $i = 1, \ldots, N$, as functions of $c^L_{t,i}$, $t = 1, \ldots, \tau+1$. We also use the notation P^{g^*} for the right-hand side in (1) which thus represents an expression in terms of the distribution of the data under the assumption that the conditional densities of $L_i(t)$, given $Pa(L_i(t))$, are constant in i as functions of $c^L_{t,i}$, indexed by the choice of stochastic intervention g^*, while one needs the causal model and randomization assumption in order to have that the right-hand side actually models the counterfactual post-intervention distribution P_{g^*}. This shows that $\psi^F_0 = \Psi(P_0)$ for a mapping Ψ from the distribution of O to the real line. Strictly speaking this does not establish a desired identifiability result yet. To start with, we need to realize that P^N_0, $\psi^{F,N}_0$, ψ^N_0 are indexed by N, and we only observed one draw from P^N_0. Therefore, we still need to show that we can construct an estimator based on O^N that is consistent for ψ^N_0 as $N \rightarrow \infty$. For that purpose, we note that the distribution P^{g^*} is identified by the common conditional distributions $P_{L(t)}$, $t = 1, \ldots, \tau + 1$, and $P_{L(0)}$ with $L(0) = (L_i(0) : i = 1, \ldots, N)$. We can construct estimators of these common conditional distributions $P_{L(t)}$ based on MLE that are consistent as $N \rightarrow \infty$, which follows from our presentation of estimators and theory. This demonstrates the identifiability of $P_{L(t)}$ as $N \rightarrow \infty$. In addition, our target parameter involves an average w.r.t. $P_{L(0)}$ which can be consistently estimated by a sample mean over $L_i(0)$, $i = 1, \ldots, N$. This demonstrates the desired identifiability of $\psi^{F,N}_0$ from the observed data as $N \rightarrow \infty$.

Likelihood and statistical model: The likelihood of the data
$O = (L(0), A(0), \ldots, L(\tau), A(\tau), Y = L(\tau + 1))$ is given by:

$$\begin{aligned} P_{Q,g}(L, A) &= \prod_{i=1}^{N} P_{L_i(0)}(L_i(0)) \\ & \prod_{i=1}^{N}\prod_{t=1}^{\tau+1} P_{L(t)}(L_i(t) \mid c^L_{t,i}))g_t(A_i(t) \mid c^A_{t,i}) \\ &= \prod_{i=1}^{N}\prod_{t=0}^{\tau+1} Q_{L(t)}(L_i(t) \mid c^L_{t,i})g_t(A_i(t) \mid c^A_{t,i}). \end{aligned}$$

We denote the factors representing the conditional distributions of $L_i(t)$ with $Q_{L(t)}$, where these conditional densities at $L_i(t)$, given $Pa(L_i(t))$, are constant in i, as functions of $L_i(t)$ and $c^L_{t,i}$. Similarly, we model the g-factor. Let $Q = (Q_t : t = 0, \ldots, \tau + 1)$ represent the collection of all these factors, and $g = (g_t : t = 0, \ldots, \tau)$, so that the distribution of O is defined by (Q, g). The conditional distributions $Q_{L(t)}(L(t) \mid c^L_t)$ are unspecified functions of $L(t)$ and c^L_t, beyond that for each value of c^L_t it is a conditional density. Similarly, the conditional distributions g_t are unspecified conditional densities. This defines now a statistical parameterization of the distribution of O in terms of Q, g, and a corresponding statistical model $\mathcal{M} = \{P_{Q,g} : Q \in \mathcal{Q}, g \in \mathcal{G}\}$, where $\mathcal{Q}$ and $\mathcal{G}$ denote the parameter spaces for Q and g, respectively.

Statistical Target Parameter: Suppose that our target parameter is $E\bar{Y}^{g^*}$ which is a function of the intervention-specific distribution P^{g^*}. Thus $E\bar{Y}^{g^*} = \Psi(P_{Q,g}) = \Psi(Q)$ depends on the distribution P of the data O through the Q-factor. Note that Q is determined by $Q_{L(0)}, \ldots, Q_{L(\tau+1)}$, i.e., the conditional distributions of $L_i(t)$, given $(\bar{A}(t-1), \bar{L}(t-1))$, which, by assumption, equal a common function $Q_{L(t)}(L_i(t) \mid c^L_{t,i})$. We can represent this statistical target parameter also as an iterative conditional expectation involving the iterative integration w.r.t. $Q_{L(t)}$, $g^*_{A(t-1)}$, starting at $t = \tau + 1$ and moving backwards till the expectation over $L(0)$:

$$\begin{aligned}
&\bar{Q}_{\tau+2} \equiv \bar{Y} \\
&\bar{Q}_{\tau+1,1} = E_{Q_{\tau+1}}(\bar{Q}_{\tau+2} \mid \bar{A}(\tau), \bar{L}(\tau)) \\
&\bar{Q}_{\tau+1} = E_{g^*_\tau}(\bar{Q}_{\tau+1,1} \mid \bar{A}(\tau-1), \bar{L}(\tau)) \\
&\text{Iterate, } t = \tau, \ldots, 0 \\
&\bar{Q}_{t+1,1} = E_{Q_{t+1}}(\bar{Q}_{t+1} \mid \bar{A}(t), \bar{L}(t)) \\
&\bar{Q}_{t+1} = E_{g^*_t}(\bar{Q}_{t+1,1} \mid \bar{A}(t-1), \bar{L}(t)) \\
&\bar{Q}_{t=0} = E_{L(0)}\bar{Q}_1 \\
&= E\bar{Y}^*
\end{aligned}$$

This representation allows the effective evaluation of $\Psi(Q)$ by first evaluating a conditional expectation w.r.t conditional distribution of $L(\tau+1)$, and thus w.r.t. $\prod_{i=1}^N Q_{L(\tau+1)}(L_i(\tau+1) \mid c^L_{\tau+1,i})$, then the conditional mean of the previous conditional expectation w.r.t. conditional distribution of $A^*(\tau)$, and iterating this process of taking a conditional expectation w.r.t. $L(t)$ and $A^*(t-1)$ till we end up with a conditional expectation over $A^*(0)$, given $L(0)$, and finally we take the marginal expectation w.r.t. the distribution of $L(0)$. Note that each conditional expectation involves an expectation over vector $(L_i(t) : i = 1, \ldots, N)$ or $(A^*_i(t-1) : i = 1, \ldots, N)$ w.r.t. product measure of common conditional distributions $Q_t(L_i(t) \mid c^L_{t,i})$ or $g^*_{t-1}(A^*_i(t-1) \mid c^{A^*}_{t-1,i})$, $t = 1, \ldots, \tau + 1$.

Statistical estimation problem: We have now defined a statistical model $\mathcal{M}$ for the distribution of O, and a statistical target parameter mapping $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ for which $\Psi(P_{Q,g})$ only depends on Q. We will also denote this target parameter with $\Psi(Q)$, with some abuse of notation by letting Ψ represent these two mappings. Given a single draw $O \sim P_{Q_0,g_0}$, we want to estimate $\Psi(Q_0)$. In addition, we want to construct an asymptotically valid confidence interval. Note that our notation suppressed the dependence on N of the data distribution $P_{Q,g}$, statistical model $\mathcal{M}$, and target parameter Ψ.

Weakening of the sequential conditional independence assumption: Consider an ordering of the units, and factorize the likelihood of the data O accordingly:

$$P^n(O) = \prod_{i=1}^N \prod_{t=0}^{\tau+1} P(L_i(t) \mid Pa(L_i(t))) \prod_{t=0}^{\tau} P(A_i(t) \mid Pa(A_i(t))),$$

where $Pa(L_i(t)) = (\bar{L}(t-1), \bar{A}(t-1), L_1(t), \ldots, L_{i-1}(t))$, and $Pa(A_i(t)) = (\bar{L}(t), \bar{A}(t-1), A_1(t), \ldots, A_{i-1}(t))$. This corresponds with the likelihood of the observed data under the causal model that does not make the sequential conditional independence assumption. However, obviously, some conditional independence structure will be needed. A reasonable assumption is that $L_i(t)$ and $L_j(t)$ are conditionally independent, given $\bar{L}(t-1), \bar{A}(t-1)$, if $F_i(t) \cap F_j(t) = \emptyset$, which corresponds with the following parent-set definitions:

$$\begin{aligned} Pa(L_i(t)) &= (\bar{L}(t-1), \bar{A}(t-1), L_j(t) : j \in \{1, \ldots, i-1\}, F_j(t) \cap F_i(t) = \emptyset) \\ Pa(A_i(t)) &= (\bar{L}(t), \bar{A}(t-1), A_j(t) : j \in \{1, \ldots, i-1\}, F_j(t) \cap F_i(t) = \emptyset). \end{aligned}$$

Such a conditional independence assumption still only allows a unit i at time t to be dependent on a limited universally bounded number of other units at time t, conditional on the past, so that our formal analysis tools as used in this article will still be applicable to establish a central limit theorem. We return to this extended semi parametric model in our sections at the end of the article dealing with the special case that $\tau = 0$.

4 Motivating examples covered by the general estimation problem studied in this article

Independent units: By defining the parent sets of $L_i(t)$ and $A_i(t)$ as only the past of unit i, the estimation problem includes estimation of a causal effect of time-dependent treatment on an outcome based on observing N independent units. If one observes multiple independent clusters of units over time, then the parent sets for unit i only include the data on units in the same cluster as unit i.

A small number of clusters with network information: If one only observes a single cluster or very small number of clusters of units, then it will be crucial for dealing with the curse of dimensionality that each unit i is only affected by a known low dimensional summary measures of the observed past on all N units: in particular, one might know the collection of current "friends" for unit i. For example, we may wish to determine if early treatment of HIV-infected patients improves the overall outcome (e.g., proportion of death) for a network of individuals, in which case the early treatment of an individual will both have a direct effect on the individual outcome, as well an indirect effect on other individuals by being less infectious or engaged in less risky sexual behaviors. In such a study one might observe the sexual or social networks of the individuals or be able to specify sets that will contain these networks.

Cluster level interventions: If one observes multiple clusters and one is interested in evaluating the effect of a cluster level intervention, then one defines $A_i(0)$ as the cluster level intervention assigned to the cluster unit i belongs to, where $A_i(0)$ is constant across i in that cluster. One can also define $L_i(0)$ as constant across units in a cluster, so that $L_i(0)$ represents cluster level baseline covariates. One might be interested in the joint effect of a cluster level intervention and a unit level exposure. In that case, one can define $L_i(0)$ as a cluster-level baseline covariate, $A_i(0)$ as a cluster-level intervention, $L_i(1)$ as unit level baseline covariates, $A_i(1)$ as unit level exposure, and $L_i(2)$ as a unit level outcome. Of course, this can be generalized to joint interventions of a single time point community level intervention and multiple time point unit level exposure. In this manner, our general formulation incorporates hierarchical two level data structures, and estimation of causal effects of joint cluster level and unit level interventions.

Current practice for assessing effects of community level interventions is heavily dominated by the use of parametric regression models, in which the coefficient in front of a treatment or exposure represents the estimand of interest. A helpful article is Oakes (2004), which reviews methods for causal inference for neighborhood effects in social epidemiology up till 2004. We refer to this article as an overview article putting this causal inference problem in context of the social epidemiology and some of the causal inference literature. Overall, from his article one concludes that the causal inference literature up till 2004 has not focussed much at all on community based interventions, and generalized mixed linear regression models, incorporating the hierarchical structure, have dominated this literature instead. The literature for assessing community level interventions involves both observational studies and randomized controlled trials. Randomized community trials involve randomly assigning an intervention among a set of possible interventions to a collection of communities/neighborhoods. Some reviews of the literature on community randomized trials are given by Donner and Klar (2000), Hayes and Moulton (2009), Campbell et al. (2007a)). Examples of randomized community trials, such as mass-media campaigns to improve health knowledge, the repair of bad sidewalks, or community policing initiatives, are provided in (Charlton et al. (1985), Meyer et al. (1991), Shipley et al. (1995), Holder et al. (1997), Feldman et al. (1998), LeFort et al. (1998), Persky et al. (1999), Biglan et al. (2000), Luepker et al. (2000)). Other examples of recent cluster randomized trials include the SPACE Study of a school level intervention to improve physical activity in Denmark(Toftager et al., 2011), a cluster randomized trial of routine HIV-1 viral load monitoring in Zambia (Koethe et al., 2010), and the PRISM trial of a community level intervention to prevent post-partum depression in Australia (Watson et al., 2004).

Examples involving the observational study of contexts such as neigborhoods are presented in Cassel (1976), McMichael (1999), Susser (1999), Berkman et al. (2000), Krieger (2001), Parsons (1951), Starr (1982), Rose (1985), Clark et al. (1991), Barr (1995), McKinlay (1996), Feldman et al. (1997)). Social epidemiology concerns the study of effects of social forces and relationships on health, and is thus a field that is primarily interested in assessing effects of community level interventions.

Oakes states, after having stressed the enormous literature on contextual effects: "Yet due largely to persistent and complex methodological obstacles, along with a lack of attention to them, the causal effect of neighborhood contexts on health continues to confuse and elude us (see Hook (2001)). There appear to be no multilevel neighborhood effect studies with observational data, including those cited above, that directly confront causal inference."

Oakes proceeds to motivate causal models for the mean counterfactual outcome of an individual under set neighborhood interventions, thereby defining a causal effect of a neighborhood intervention on an individual outcome. He presents mixed linear models for the counterfactual mean outcome as a function of an individual and neighborhood specific covariates. He considers the required randomization assumption and experimental treatment assignment assumption, well known in the causal inference literature, under which the coefficients in the mixed linear model can be interpreted as a conditional causal effect, within strata of the covariates that entered the model.

Oakes presents the following comment on the enormous use (and abuse) of mixed linear models. We quote from Oakes review "The theoretical foundation of multilevel models lies in variance component methodology, which in its modern form dates back to Fishers work circa 1925 (Draper (1995)). A ground-breaking advance came when Lindley and Smith (1972) formulated their empirical Bayes regression model, but it was not until the introduction of the EM algorithm (Dempster et al. (1977)) that computational feasibility was obtained. Laird and Ware (1982) popularized the model for biostatisticians, Bryk, Raudenbush, Goldstein and Mason for social scientists (Mason et al. (1984), Goldstein (1987), Bryk and Raudenbush (1992)). From our perspective, the widespread (ab)use of the model is due to the recent introduction of user- friendly software, especially HLM and MlWin, and an accessible translation for SAS users by Verbeke and Molenbergs (1997) and Singer (1998). See also Kreft et al. (1994), Leeuw and Kreft (2001). ”

Oakes also states: ”Understandably, none of the more recent and rigorous discussions of causal inference in either epidemiology or social science (Susser (1973),Greenland (1990), Greenland (2001), Greenland (2002), Manski (1993), Halloran and Struchiner (1995), Morgenstern (1995), M.E.Sobel (1995), Kaufman and Cooper (1995), Kaufman and Poole (2000), Kaufman and Kaufman (2001), Robins (2001), Maldonado and Greenland (2002)) addressed multilevel neighborhood effects research directly. Finally, none of the many noteworthy general discussions on causal inference with observational data (e.g. Campbell and Stanley (1963), Cochran (1965), McKinlay (1975), Heckman (1979), Leamer (1983), Smith (1990), Rubin (1991), Clogg and Haritou (1997), Copas and Li (1997), Freedman (1997), Winship and Morgan (1999), Pearl (2000), Rosenbaum (2002)) address neighborhood effects or multilevel mod- els, which appear to present some unique issues.”

Multiple connected clusters: Suppose that one observes multiple clusters of units, but that neighboring clusters interfere with each other. As above, one can define $L_i(0)$ as a cluster level covariate, $A_i(0)$ as a cluster level treatment, and $L_i(t)$, $A_i(t)$, as unit level covariates and exposure, $t = 1, \ldots, \tau$, and $Y_i = L_i(\tau + 1)$ as the unit level outcome of interest. In this case, the parents of the unit level exposures $A_i(t)$ and covariates $L_i(t)$ can now include units from neighboring clusters as well. If no network information is available, these friends would be defined as all units in the neighboring clusters, including its own cluster. The cluster level treatment $A_i(0)$ could not only depend on $L_i(0)$, but also be a function of the cluster level covariate values of the neighboring clusters. Again, to deal with the curse of dimensionality, one will either need to observe many clusters, or one needs network information at the unit level.

Group sequential adaptive designs involving adaptive randomization: Suppose that a subject i enters the study at a random time t_i, baseline covariates are collected, the subject is subsequently randomized to a treatment, and is then followed up for K time-points at which time point an outcome is collected. Suppose that the probability of receiving one of the available treatments is a function of the data measured on any of the subjects that has entered the study before time t_i, while baseline covariates and outcomes are generated independently across the subjects. This includes a group sequential design with adaptive randomization of treatment as analyzed in van der Laan (2008); Chambaz and van der Laan (2010b, 2011a,b), Bai et al. (2002); Andersen et al. (1994); Flournoy and Rosenberger (1995); Hu and Rosenberger (2000); Rosenberger (1996); Rosenberger et al. (1997); Rosenberger and Grill (1997); Rosenberger and Shiram (1997); Tamura et al. (1994); Wei (1979); Wei and Durham (1978); Wei et al. (1990); Zelen (1969); Cheng and Shen (2005); van der Laan (2008); Chambaz and van der Laan (2010a); van der Laan and Rose (2012) In this case we define $\tau + 1$ as the total length of the study from the start $t = 0$ of the study at which the first subject entered. At each time point $0 \leq t < t_i$, we can define $A_i(t) = 0$, and at $t = t_i$ $A_i(t)$ is defined as the treatment A_i the subject received, while for $t > t_i$, $A_i(t) = A_i$ is degenerate. At each time point $0 \leq t < t_i$, $L_i(t)$ is defined as the process $I(t_i \leq t)$, at $t = t_i$, $L_i(t_i)$ includes the baseline covariates W_i and it also includes observing that t_i occurred, at $t_i < t < t_i + K$, $L_i(t)$ is defined constant, at time $t = t_i + K$, $L_i(t)$ is defined as the outcome Y_i, and $L_i(t)$ remains constant for $t_i + K < t < \tau + 1$. This defines now $L_i(0), A_i(0), L_i(1), \ldots, L_i(\tau), A_i(\tau), L_i(\tau + 1)$. Our intervention on the time-dependent treatment $A_i(t)$ is a dynamic intervention that only intervenes on the A_i-nodes at time t_i at which it sets the value of A_i equal to a specific treatment. This is indeed a dynamic intervention that responds to the indicator process $I(t_i \leq t) \subset L_i(t)$ that jumps when t_i occurs. Our model allows us to extend the group sequential adaptive designs to allow for outcomes of the subjects to be affected by what has been observed on previously recruited subjects.

5 Maximum Likelihood Estimation, Cross-Validation, Super-Learning, and Targeted Maximum Likelihood Estimation

We could estimate the distribution of $L(0)$ with the empirical distribution that puts mass 1 on $(L_i(0) : i = 1, \ldots, N)$. In general, we estimate the distribution of $L(0)$ with the NPMLE that maximizes the log-likelihood $\sum_i \log Q_{L_i(0)}(L_i(0))$ over all possible distributions of $L(0)$ that the statistical model $\mathcal{M}$ allows. In particular, if it is known that $L_i(0)$ are i.i.d., then we would estimate the common distribution Q_0 of $L_i(0)$ with the empirical distribution that puts mass $1/N$ on $L_i(0)$, $i = 1, \ldots, N$.

Regarding estimation of $Q_t = Q_{L(t)}$ for $t = 1, \ldots, \tau + 1$, we consider the log-likelihood loss function for Q_t:

$$L_t(Q) = -\sum_{i=1}^{N} \log Q_t(L_i(t) \mid c^L_{i,t}).$$

We can use loss-based cross-validation and super-learning to fit this function Q_t of $(l(t), c^L_t)$, utilizing that $L_i(t)$, $i = 1, \ldots, N$, are conditionally independent, given $\bar{A}(t-1), \bar{L}(t-1)$. If $L_i(t)$ is continuous and/or multi-dimensional, one could code it in terms of binary variables $I(L_i(t) = l)$, and model the conditional distribution/hazard of $I(L_i(t) = l)$, given $L_i(t) \geq l$ and $\bar{A}(t-1), \bar{L}(t-1)$, as a function of $c^L_{t,i}$ and l, as in van der Laan (2010a,b). This allows one to utilize estimators of predictors of binary variables in the machine learning literature, including standard logistic regression software for fitting parametric models.

We could fit each Q_t separately for $t = 1, \ldots, \tau + 1$, but it is also possible to pool across t based on the sum loss function

$$L(Q) = \sum_t L_t(Q).$$

Similarly, we can use the log-likelihood loss-function for g_t:

$$L_t(g) = \sum_{i=1}^{N} \log g_t(A_i(t) \mid c^A_{t,i}),$$

and use loss-based cross-validation and super-learning to fit g_t.

Given the resulting estimator Q_N of Q_0, one can evaluate $\Psi(Q_N)$ as estimator of $\psi_0 = \Psi(Q_0)$, according to the iterative conditional expectation mapping presented earlier. Since Q_N is optimized to fit Q_0, such a plug-in estimator is overly biased for $\Psi(Q_0)$.

Targeted Maximum Likelihood Estimation (TMLE): TMLE will involve modifying $Q_{t,N}$ into a targeted version $Q^*_{t,N}$, $t = 0, \ldots, \tau + 1$, through utilization of an estimator g_n of g_0, and a least favorable submodel through Q_N, and the resulting estimator of ψ_0 is defined accordingly as $\Psi(Q^*_N)$. Thus, a TMLE will also involve estimation of the intervention mechanism $g_0 = (g_{0,t} : t = 0, \ldots, \tau)$. To define and understand the asymptotic behavior of such a TMLE we need to determine the efficient influence curve of the statistical target parameter, and understand the expectation of the efficient influence curve under misspecified nuisance parameters. We provide the TMLE in detail in Section 8.

6 Efficient influence curve of target parameter

In order to define a targeted maximum likelihood estimator we derive the efficient influence curve of $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ defined on our model $\mathcal{M}$ in which the distribution of O is parameterized by $Q = (Q_t : t = 0, \ldots, \tau + 1)$ and $g = (g_t : t = 0, \ldots, \tau)$. In order to solve this problem we first determine the efficient influence curve for other general statistical models implied by statistical graphs, before tackling our actual model that has additional structure due to having common conditional distributions across i.

6.1 Efficient influence curve and its robustness for general nonparametric statistical graphs.

We start out with determining the efficient influence curve for a general nonparametric statistical graph.

Theorem 1 *Consider longitudinal data structure* $O = (L(0), A(0), \ldots, L(\tau), A(\tau), Y = L(\tau + 1)) \sim P$. *Let* $g = \prod_{t=0}^{\tau} g_t$, $g_t = P(A(t) \mid \bar{L}(t), \bar{A}(t-1))$, $Q = \prod_{t=0}^{\tau+1} Q_t$, $Q_t = P(L(t) \mid \bar{L}(t-1), \bar{A}(t-1))$, *so that the density* P *factorizes as* $P = Qg$. *Consider the statistical model* $\mathcal{M}$ *that assumes* $g_t(A(t) \mid \bar{A}(t-1), \bar{L}(t)) = g_t(A(t) \mid c_t^A)$, *and* $Q_t(L(t) \mid \bar{L}(t-1), \bar{A}(t-1)) = Q_t(L(t) \mid c_t^L)$ *for some specified functions* $C_t^A = c_t^A(\bar{A}(t-1), \bar{L}(t))$ *and* $C_t^L = c_t^L(\bar{A}(t-1), \bar{L}(t-1))$, *respectively,* $t = 1, \ldots, \tau + 1$. *For notational convenience, let* c_0^L *be a constant so that conditioning on this can be ignored. Let* $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ *be defined as* $E_{Q,g^*}Y$, *where* $P_{Q,g^*} = Qg^*$ *is the G-computation formula for the distribution of* O *under stochastic intervention* g^*. *The efficient influence curve* $D^*(Q, g)$ *of* Ψ *at* $P_{Q,g}$ *is given by*

$$D^*(Q,g) = \sum_{t=0}^{\tau+1} \left\{ E_{Q,g}(Yg^*/g \mid L(t), C_t^L) - E_{Q,g}(Yg^*/g \mid C_t^L) \right\}.$$

Define

$$h_{t,Q,g}(c_t) = P_{Q,g}(C_t^L = c_t) \text{ and } h_{t,Q,g^*}(c_t) = P_{Q,g^*}(C_t^L = c_t).$$

For $t = 0$, *we define* $h_{t,Q,g} = h_{t,Q,g^*} = 1$. *The efficient influence curve can be represented as:*

$$\begin{aligned}
&D^*(Q,g)(L(t), C_t^L) = \\
&\textstyle\sum_{t=0}^{\tau+1} \int E_{Q,g^*}(Y \mid L(t), \bar{A}(t-1), \bar{L}(t-1)) \frac{I(C_t^L(\bar{A}(t-1), \bar{L}(t-1)) = c_t^L)}{h_{t,Q,g}(c_t^L)} \prod_{l=0}^{t-1} g_l^* Q_l \\
&- \textstyle\sum_{t=0}^{\tau+1} \int E_{Q,g^*}(Y \mid \bar{A}(t-1), \bar{L}(t-1)) \frac{I(C_t^L(\bar{A}(t-1), \bar{L}(t-1)) = c_t^L)}{h_{t,Q,g}(c_t^L)} \prod_{l=0}^{t-1} g_l^* Q_l \\
&= \textstyle\sum_{t=0}^{\tau+1} \left\{ E_{Q,g^*}(Y \mid L(t), C_t^L) - E_{Q,g^*}(Y \mid C_t^L) \right\} \frac{h_{t,Q,g^*}(C_t^L)}{h_{t,Q,g}(C_t^L)}.
\end{aligned}$$

Thus, if $E_{Q,g^*}(Y \mid L(t), \bar{A}(t-1), \bar{L}(t-1))$ *only depends on* $L(t), C_t^L$*, then this reduces to*

$$D^*(Q,g) = \textstyle\sum_{t=0}^{\tau+1} \frac{h_{t,Q,g^*}(C_t^L)}{h_{t,Q,g}(C_t^L)}$$
$$\left\{E_{Q,g^*}(Y \mid L(t), \bar{A}(t-1), \bar{L}(t-1)) - E_{Q,g^*}(Y \mid \bar{A}(t-1), \bar{L}(t-1))\right\}.$$

Robustness of efficient influence curve: *Let* $P_0 = Q_0 g_0$*. If* $h_{t,Q,g} = h_{t,Q_0,g_0}$ *and* $Q_t = Q_{0,t}$*, then* $P_0 D_t^*(Q,g) = 0$*,* $t = 0, \ldots, \tau+1$*. Also, if* $h_{t,Q,g} = h_{t,Q_0,g_0}$ *and* $P_0 D_t^*(Q,g) = 0$*, then* $E_{Q_{-t},Q_{0,t},g^*}Y - E_{Q,g^*}Y = 0$*, where we use notation* $Q_{-t} = (Q_s : s \neq t)$*.*

If $h_{t,Q,g} = h_{t,Q_0,g_0}$ *for all* t*, then*

$$P_0 D^*(Q,g) = \sum_{t=0}^{\tau+1} \{E_{Q_{-t},Q_{0,t},g^*}Y - E_{Q,g^*}Y\}.$$

Define

$$R(Q,Q_0) \equiv \sum_{t=0}^{\tau+1} E_{Q_{0:t-1}-Q_{0:0:t-1},Q_{0,t}-Q_t,Q_{t+1:\tau+1},g^*}Y,$$

where we used notation $Q_{l:m} = (Q_s : l \leq s \leq m)$*. Note that* $R(Q,Q_0)$ *is a second order term involving square differences between* Q *and* Q_0*. This implies that, if* $h_{t,Q,g} = h_{t,Q_0,g_0}$ *for all* t*, then*

$$P_0 D^*(Q,g) = \Psi(Q_0) - \Psi(Q) + R(Q_0,Q).$$

Define

$$R_t((h_t,h_{t,0}),(Q,Q_0))$$
$$= P_0\left\{E_{Q,g^*}(Y \mid L(t), C_t^L) - E_{Q,g^*}(Y \mid C_t^L)\right\} \frac{h_{t,Q,g^*}(h_{t,Q,g} - h_{t,Q_0,g_0})}{h_{t,Q,g} h_{t,Q_0,g_0}}.$$

Since the conditional expectation w.r.t. $Q_{0,t}$ *of the difference of the two conditional expectations equals zero if* $Q_t = Q_{0,t}$*, it follows that the right-hand side also involves a difference* $Q_{0,t} - Q_t$*, and thereby a product* $(Q_{0,t} - Q_t)(h_{t,Q,g} - h_{t,Q_0,g_0})$*. Let* $R((h,h_0),(Q,Q_0)) = \sum_{t=0}^{\tau+1} R_t((h_t,h_{t,0}),(Q,Q_0))$*. In general, we have*

$$P_0 D^*(Q,g) = \Psi(Q_0) - \Psi(Q) + R((h,h_0),(Q,Q_0)) + R(Q,Q_0).$$

In particular, for all g*,* $P_0 D^*(Q_0,g) = 0$*.*

Proof. The efficient influence curve is not affected by the choice of model on g, since the target parameter is only a function of the Q-factor, and the likelihood factorizes as $P_{Q,g} = Qg$. The efficient influence curve is given by $\sum_t E(Yg^*/g \mid L(t), C_t^L) - E(Yg^*/g \mid C_t^L)$. Firstly, we use that

$$E(Yg^*/g \mid L(t), C_t^L) = E(E(Yg^*/g \mid L(t), \bar{A}(t-1), \bar{L}(t-1)) \mid L(t), C_t^L).$$

The inner conditional expectation is given by

$$\begin{aligned}
&\textstyle\int Y \frac{\prod_{l=0}^{\tau} g_l^*}{\prod_{l=0}^{\tau} g_l} \prod_{l=t}^{\tau+1} g_l \prod_{l=t+1}^{\tau+1} Q_l = \int Y \prod_{l=0}^{t-1} g_l^*/g_l \prod_{l=t}^{\tau+1} g_l^* \prod_{l=t+1}^{\tau+1} Q_l \\
&\textstyle= \prod_{l=0}^{t-1} g_l^*/g_l E_{Q,g^*}(Y \mid L(t), \bar{A}(t-1), \bar{L}(t-1)).
\end{aligned}$$

The conditional distribution of $(L(t), \bar{A}(t-1), \bar{L}(t-1))$, given $L(t), C_t^L = c_t^L$, can be written as

$$\begin{aligned}
&I(c_t^L(\bar{A}(t-1), \bar{L}(t-1)) = c_t^L)\tfrac{P(\bar{L}(t), \bar{A}(t-1))}{P(L(t), c_t^L)} \\
&= I(c_t^L(\bar{A}(t-1), \bar{L}(t-1)) = c_t^L)P(\bar{L}(t-1), \bar{A}(t-1))/P(C_t^L = c_t^L) \\
&= I(c_t^L(\bar{A}(t-1), \bar{L}(t-1)) = c_t^L)\tfrac{\prod_{l=0}^{t-1} g_l \prod_{l=0}^{t-1} Q_l}{h_{t,Q,g}(c_t^L)}.
\end{aligned}$$

Let $Pa(L(t)) = (\bar{A}(t-1), \bar{L}(t-1))$. This proves that

$$\begin{aligned}
&E_{Q,g}(Yg^*/g \mid L(t), C_t^L = c_t^L) \\
&\textstyle= \int E_{Q,g^*}(Y \mid L(t), Pa(L(t)))I(c_t^L(Pa(L(t))) = c_t^L)\tfrac{\prod_{l=0}^{t-1} g_l^* \prod_{l=0}^{t-1} Q_l}{h_{t,Q,g}(c_t^L)} \\
&\textstyle= \int E_{Q,g^*}(Y \mid L(t), C_t^L = c_t^L)\tfrac{h_{t,Q,g^*}(c_t^L)}{h_{t,Q,g}(c_t^L)},
\end{aligned}$$

where we integrated over all $(\bar{A}(t-1), \bar{L}(t-1))$ for which $C_t^L = c_t^L$ to obtain the last expression. This proves the representation of the efficient influence curve (2). The representation under the condition that $E_{Q,g^*}(Y \mid L(t), \bar{A}(t-1), \bar{L}(t-1))$ only depends on $L(t), c_t^L$ is an immediate consequence.

We now wish to establish the expectation of efficient influence curve w.r.t. P_0. The expectation w.r.t. P_0 of $D_t^*(Q, g)$ corresponds with integration w.r.t. $Q_{t,0}h_{t,Q_0,g_0}$. So if $h_{t,Q,g} = h_{t,Q_0,g_0}$, and $h_{t,Q,g^*} = h_{t,Q_0,g^*}$, then we have

$$\begin{aligned}
&P_0 E_{Q,g}(Yg^*/g \mid L(t), c_t^L) \\
&\textstyle= \int_{c_t} \int_{L(t)} \int_{\bar{A}(t-1),\bar{L}(t-1)} E_{Q,g^*}(Y \mid L(t), Pa(L(t)))I(c_t^L(Pa(L(t))) = c_t) \\
&\qquad\qquad\textstyle\prod_{l=0}^{t-1} g_l^* \prod_{l=0}^{t-1} Q_l Q_{0,t}(L(t) \mid c_t^L) \\
&\textstyle= \int_{L(t)} \int_{\bar{A}(t-1),\bar{L}(t-1)} E_{Q,g^*}(Y \mid L(t), Pa(L(t))) \\
&\qquad\qquad\textstyle\prod_{l=0}^{t-1} g_l^* \prod_{l=0}^{t-1} Q_l Q_{0,t}(L(t) \mid c_t^L(Pa(L(t))) \\
&= E_{Q_{-t},Q_{0,t},g^*} Y,
\end{aligned}$$

where we used notation $Q_{-t} = (Q_l : l \neq t)$. Note that we used Fubini's theorem to first integrate out c_t, using that the indicator will only be non-zero at $c_t = c_t^L(Pa(L(t)))$. As a consequence of the disappearing of the indicator, we end up with a pure marginal expectation w.r.t. $P_{Q_{-t},Q_{0,t},g^*}Y$. Similarly, we obtain at a Q with $h_{t,Q,g} = h_{t,Q_0,g_0}$,

$$P_0 E_{Q,g}(Yg^*/g \mid L(t), c_t^L) = E_{Q,g^*}Y.$$

Thus,

$$P_0 D_t^*(Q,g) = E_{Q_{-t},Q_{0,t},g^*}Y - E_{Q,g^*}Y.$$

Thus, this proves that if $h_{t,Q,g} = h_{t,Q_0,g_0}$ and $Q_t = Q_{0,t}$, then $P_0 D_t^*(Q,g) = 0$, $t = 0, \ldots, \tau+1$. Or, equivalently, if $h_{t,Q,g} = h_{t,Q_0,g_0}$ and $P_0 D_t^*(Q,g) = 0$, then $E_{Q_{-t},Q_{0,t},g^*}Y - E_{Q,g^*}Y = 0$.

If $h_{t,Q,g} = h_{t,Q_0,g_0}$ for all t, then

$$P_0 D^*(Q,g) = \sum_{t=0}^{\tau+1} \{E_{Q_{-t},Q_{0,t},g^*}Y - E_{Q,g^*}Y\}.$$

On the other hand, we have

$$\Psi(Q_0) - \Psi(Q) = \sum_{t=0}^{\tau+1} E_{Q_{0,0:t-1},Q_{0,t},Q_{t+1:\tau+1},g^*}Y - E_{Q_{0,0:t-1},Q_t,Q_{t+1:\tau+1},g^*}Y.$$

We have

$$\begin{array}{l}
\sum_{t=0}^{\tau+1}\{E_{Q_{-t},Q_{0,t},g^*}Y - E_{Q,g^*}Y\} - \\
\sum_{t=0}^{\tau+1}\{E_{Q_{0,0:t-1},Q_{0,t},Q_{t+1:\tau+1},g^*}Y - E_{Q_{0,0:t-1},Q_t,Q_{t+1:\tau+1},g^*}Y\} \\
= \sum_{t=0}^{\tau+1} E_{Q_{0:t-1}-Q_{0:0:t-1},Q_{0,t}-Q_t,Q_{t+1:\tau+1},g^*}Y \\
\equiv R(Q_0,Q)
\end{array}$$

We note that $R(Q,Q_0)$ is a second order term involving square differences between Q and Q_0. This proves that if $h_{t,Q,g} = h_{t,Q_0,g_0}$ for all t, then

$$P_0 D^*(Q,g) = \Psi(Q_0) - \Psi(Q) + R(Q_0,Q).$$

It also follows that $P_0 D^*(Q,g) = R((h,h_0),(Q,Q_0)) + R(Q,Q_0) + \Psi(Q_0) - \Psi(Q)$, where $R((h,h_0),(Q,Q_0))$ is involves product differences of $(h-h_0)$ and $Q - Q_0$, which is thus also a second order term. Specifically,

$$R_t((h_t,h_{t,0}),(Q,Q_0)) = P_0\left\{E_{Q,g^*}(Y \mid L(t), C_t^L) - E_{Q,g^*}(Y \mid C_t^L)\right\} \frac{h_{t,Q,g^*}(h_{t,Q,g} - h_{t,Q_0,g_0})}{h_{t,Q,g}h_{t,Q_0,g_0}}.$$

Since the conditional expectation w.r.t. $Q_{0,t}$ of the difference of the two conditional expectations equals zero if $Q_t = Q_{0,t}$, it follows that the right-hand side also involves a difference $Q_{0,t} - Q_t$, and thereby a product $(Q_{0,t} - Q_t)(h_{t,Q,g} - h_{t,Q_0,g_0})$.

This proof also shows that $P_0 D^*(Q_0, g) = 0$ for all g. More explicitly,

$$\begin{array}{l} P_0 E_{Q_0,g}(Yg^*/g \mid L(t), c_t^L) \\ = \int_{c_t} \int_{L(t)} \int_{\bar{A}(t-1),\bar{L}(t-1)} E_{Q_0,g^*}(Y \mid L(t), Pa(L(t))) I(c_t^L(Pa(L(t))) = c_t) \\ \prod_{l=0}^{t-1} g_l^* \prod_{l=0}^{t-1} Q_{l,0} Q_{0,t}(L(t) \mid c_t) \frac{h_{t,0}(c_t)}{h_{t,Q_0,g}(c_t)}. \end{array}$$

We now use Fubini's theorem to first integrate over $L(t)$ w.r.t. $Q_{0,t}$. Since only the conditional expectation depends on $L(t)$, this yields

$$\begin{array}{l} P_0 E_{Q_0,g}(Yg^*/g \mid L(t), c_t^L) \\ = \int_{c_t} \int_{\bar{A}(t-1),\bar{L}(t-1)} E_{Q_0,g^*}(Y \mid Pa(L(t))) I(c_t^L(Pa(L(t))) = c_t) \\ \qquad \qquad \qquad \qquad \prod_{l=0}^{t-1} g_l^* \prod_{l=0}^{t-1} Q_{l,0} \frac{h_{t,0}(c_t)}{h_{t,Q_0,g}(c_t)}. \end{array}$$

Thus, this provides another explicit proof that $P_0 D_t^*(Q_0, g) = 0$. This completes the proof of the theorem. □

The following theorem extends this result in a straightforward manner to determine the representation of the efficient influence curve corresponding with the orthogonal decomposition implied by a more general factorized likelihood. This theorem still assumes that the different factors of the likelihood are variation independent and nonparametrically modeled.

Theorem 2 *Consider longitudinal data structure* $O = (L(0), A(0), \ldots, L(\tau), A(\tau), Y = L(\tau+1) \sim P$, *and let* $L(t) = (L(t,j) : j = 1, \ldots, n_t)$, $t = 1, \ldots, \tau + 1$, *while* $L(0)$ *is not further decomposed. For notational convenience, let* $n_0 = 1$, *and* $L(0,1) = L(0)$. *Let* $g = \prod_{t=0}^{\tau} g_t$, $g_t = P(A(t) \mid \bar{L}(t), \bar{A}(t-1))$, $Q = \prod_{t=0}^{\tau+1} \prod_{j=1}^{n_t} Q_{t,j}$, $Q_{t,j} = P(L(t,j) \mid L(t,1), \ldots, L(t,j-1), \bar{L}(t-1), \bar{A}(t-1))$, *so that* $P = Qg$. *Consider the statistical model* $\mathcal{M}$ *that assumes a model on* $g_t(A(t) \mid \bar{A}(t-1), \bar{L}(t))$ *(such as* $= g_t(A(t) \mid c_t^A)$*), and*

$$Q_{t,j}(L(t,j) \mid L(t:1:j-1), \bar{L}(t-1), \bar{A}(t-1)) = Q_{t,j}(L(t,j) \mid c_{t,j}^L)$$

for some specified functions $c_t^A = c_t^A(\bar{A}(t-1), \bar{L}(t))$ *and* $c_{t,j}^L(L(t:1:j-1), \bar{A}(t-1), \bar{L}(t-1))$, *respectively,* $j = 1, \ldots, n_t, t = 1, \ldots, \tau + 1$. *The same results below apply for any statistical model on* g.

For notational convenience, let $c^{L(0)}$ *be a constant so that conditioning on this can be ignored. Let* $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ *be defined as* $E_{Q,g^*}Y$, *where* $P_{Q,g^*} =$

Qg^* is the G-computation formula for the distribution of O under stochastic intervention g^*. The efficient influence curve $D^*(Q,g)$ of $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ at $P_{Q,g}$ is given by

$$D^*(Q,g) = \sum_{t=0}^{\tau+1} \sum_{j=1}^{n_t} \left\{ E_{Q,g}(Yg^*/g \mid L(t,j), c_{t,j}^L) - E_{Q,g}(Yg^*/g \mid c_{t,j}^L) \right\}.$$

Let $\mathcal{S}(t,j) = \{(s,k) : s \in \{0,\ldots,t-1\}, k\} \cup \{(t,k) : k = 1,\ldots,j-1\}$. Define

$$h_{t,j,Q,g}(c_{t,j}) = P_{Q,g}(C_{t,j}^L = c_{t,j}) \text{ and } h_{t,j,Q,g^*}(c_{t,j}) = P_{Q,g^*}(C_{t,j}^L = c_{t,j}).$$

If $t=0$, we define $h_{t=0,Q,g} = h_{t=0,Q,g^*} = 1$.

The efficient influence curve can be represented as:

$$\begin{array}{l} D^*(Q,g) = \sum_{t=0}^{\tau+1} \sum_{j=1}^{n_t} \int E_{Q,g^*}(Y \mid L(t,j), L(t,1:j-1), \bar{A}(t-1), \bar{L}(t-1)) \\ \qquad \frac{I(C^{k,j}(\bar{A}(t-1),\bar{L}(t-1),L(t,1:j-1))=C_{t,j}^L)}{h_{t,j,Q,g}(C_{t,j}^L)} \prod_{l=0}^{k-1} g_l^* \prod_{l \in \mathcal{S}(t,j)} Q_l \\ - \sum_{t=0}^{\tau+1} \sum_{j=1}^{n_t} \int E_{Q,g^*}(Y \mid L(t,1:j-1), \bar{A}(t-1), \bar{L}(t-1)) \\ \qquad \frac{I(C^{k,j}(\bar{A}(t-1),\bar{L}(t-1),L(t,1:j-1))=C_{t,j}^L)}{h_{t,j,Q,g}(C_{t,j}^L)} \prod_{l=0}^{k-1} g_l^* \prod_{l \in \mathcal{S}(t,j)} Q_l \\ = \sum_{t=0}^{\tau+1} \sum_{j=1}^{n_t} \frac{h_{t,j,Q,g^*}(C_{t,j}^L)}{h_{t,j,Q,g}(C_{t,j}^L)} \left\{ E_{Q,g^*}(Y \mid L(t,j), C_{t,j}^L) - E_{Q,g^*}(Y \mid C_{t,j}^L) \right\}. \end{array}$$

In particular, if $E_{Q,g^*}(Y \mid L(t,j), L(t,1:j-1), \bar{A}(t-1), \bar{L}(t-1))$ only depends on $L(t,j), C_{t,j}^L$, then

$$\begin{array}{l} D^*(Q,g) = \sum_{t=0}^{\tau+1} \sum_{j=1}^{n_t} \frac{h_{t,j,Q,g^*}(C_{t,j}^L)}{h_{t,j,Q,g}(C_{t,j}^L)} \\ \left\{ E_{Q,g^*}(Y \mid L(t,j), Pa(L(t,j))) - E_{Q,g^*}(Y \mid Pa(L(t,j))) \right\}. \end{array}$$

where we used the notation $Pa(L(t,j)) = (L(t,1:j-1), \bar{A}(t-1), \bar{L}(t-1))$.
Robustness of efficient influence curve: *Let $P_0 = Q_0 g_0$. If $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$ and $Q_{t,j} = Q_{0,t,j}$, then $P_0 D_{t,j}^*(Q,g) = 0$. Also, if $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$ and $P_0 D_t^*(Q,g) = 0$, then $E_{Q_{-(t,j)},Q_{0,(t,j)},g^*} Y - E_{Q,g^*} Y = 0$.*

If $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$ for all t,j, then

$$P_0 D^*(Q,g) = \sum_{t=0}^{\tau+1} \sum_{j=1}^{n_t} \{ E_{Q_{-(t,j)},Q_{0,(t,j)},g^*} Y - E_{Q,g^*} Y \}.$$

Define

$$R(Q,Q_0) \equiv \sum_{t=0}^{\tau+1} \sum_{j=1}^{n_t} \{ E_{Q_{(t,j)-} - Q_{0,(t,j)-}, Q_{0,t,j} - Q_{t,j}, Q_{(t,j)+}, g^*} Y,$$

where $Q_{(t,j)-}$ represents all factors before (t,j) in the ordering (by column so that time-ordering is respected), and $Q_{(t,j)+}$ represents all factors after (t,j) in the ordering. Note that $R(Q,Q_0)$ is a second order term involving square differences between Q and Q_0. This implies that, if $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$ for all t,j, then

$$P_0D^*(Q,g) = \Psi(Q_0) - \Psi(Q) + R(Q_0,Q).$$

Define

$$\begin{aligned}&R_{t,j}((h_{t,j},h_{t,j,0}),(Q,Q_0))\\&= P_0\left\{E_{Q,g^*}(Y \mid L(t,j),C^L_{t,j}) - E_{Q,g^*}(Y \mid C^L_{t,j})\right\}\frac{h_{t,j,Q,g^*}(h_{t,j,Q,g}-h_{t,j,Q_0,g_0})}{h_{t,j,Q,g}h_{t,j,Q_0,g_0}}.\end{aligned}$$

Since the conditional expectation w.r.t. $Q_{0,t,j}$ of the difference of the two conditional expectations equals zero if $Q_{t,j} = Q_{0,t,j}$, it follows that the right-hand side also involves a difference $Q_{0,t,j} - Q_{t,j}$, and thereby a product $(Q_{0,t,j} - Q_{t,j})(h_{t,jQ,g} - h_{t,j,Q_0,g_0})$. Let $R((h,h_0),(Q,Q_0)) = \sum_t \sum_j R_{t,j}((h_t,h_{t,0}),(Q,Q_0))$. In general, we have

$$P_0D^*(Q,g) = \Psi(Q_0) - \Psi(Q) + R((h,h_0),(Q,Q_0)) + R(Q,Q_0).$$

In particular, for all g, $P_0D^*(Q_0,g) = 0$.

Proof. The efficient influence curve is not affected by the choice of model on g, since the target parameter is only a function of the Q-factor, and the likelihood factorizes as $P_{Q,g} = Qg$. The efficient influence curve is given by $\sum_{t,j} E(Yg^*/g \mid L(t,j),C^L_{t,j}) - E(Yg^*/g \mid C^L_{t,j})$. Firstly, we use that

$$\begin{aligned}&E(Yg^*/g \mid L(t,j),C^L_{t,j})\\&= E(E(Yg^*/g \mid L(t,j),L(t,1:j-1),\bar{A}(t-1),\bar{L}(t-1)) \mid L(t,j),C^L_{t,j}).\end{aligned}$$

The inner conditional expectation is given by

$$\begin{aligned}&\int Y\frac{\prod_{l=0}^{\tau} g^*_l}{\prod_{l=0}^{\tau} g_l}\prod_{l=t}^{\tau+1} g_l \prod_{l\in\mathcal{S}(t,j)^c,l\neq(t,j)} Q_l\\&= \int Y\prod_{l=0}^{t-1} g^*_l/g_l \prod_{l=t}^{\tau+1} g^*_l \prod_{l\in S(t,j)^c,l\neq(t,j)} Q_l\\&= \prod_{l=0}^{t-1} g^*_l/g_l E_{Q,g^*}(Y \mid L(t,j),L(t,1:j-1),\bar{A}(t-1),\bar{L}(t-1)).\end{aligned}$$

The conditional distribution of $(L(t,j),L(t,1:j-1),\bar{A}(t-1),\bar{L}(t-1),c^L_{t,j})$, given $L(t,j),C^L_{t,j} = c^L_{t,j}$, can be written as

$$\begin{aligned}&I(c^L_{t,j}(\bar{A}(t-1),\bar{L}(t-1),L(t,1:j-1)) = c^L_{t,j})\frac{P(L(t,1:j),\bar{L}(t-1),\bar{A}(t-1))}{P(L(t,j),c^L_{t,j})}\\&= I(c^L_{t,j}(\bar{A}(t-1),\bar{L}(t-1),L(t,1:j-1)) = c^L_{t,j})\frac{P(L(t,1:j-1),\bar{L}(t-1),\bar{A}(t-1))}{P(c^L_{t,j}=c^L_{t,j})}\\&= I(c^L_{t,j}(\bar{A}(t-1),\bar{L}(t-1),L(t,1:j-1)) = c^L_{t,j})\frac{\prod_{l=0}^{k-1} g_l \prod_{l\in S(t,j)} Q_l}{h_{t,j,Q,g}(c^L_{t,j})}.\end{aligned}$$

This proves that

$$\begin{aligned}
&E_{Q,g}(Yg^*/g \mid L(t,j), C^L_{t,j} = c^L_{t,j}) \\
&= \int E_{Q,g^*}(Y \mid L(t,j), Pa(L(t,j)))I(c^L_{t,j}(Pa(L(t,j))) = c^L_{t,j}) \\
&\frac{\prod_{l=0}^{k-1} g^*_l \prod_{l\in\mathcal{S}(t,j)} Q_l}{h_{t,j,Q,g}(c_{t,j})}.
\end{aligned}$$

Further integration over $Pa(L(t,j))$ satisfying $c^L_{t,j}(Pa(L(t,j)) = c^L_{t,j}$ yields the last representation, and thereby the proof of the representation of the efficient influence curve (2).

We now wish to establish the robustness w.r.t. P_0. Let $\mathcal{S}(t,j)^+ = \mathcal{S}(t,j) \cup \{t,j\}$. The expectation w.r.t. P_0 of $D^*_{t,j}(Q,g)$ corresponds with integration w.r.t. $Q_{t,j,0}h_{t,j,Q_0,g_0}$. So if $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$, we have

$$\begin{aligned}
&P_0E_{Q,g}(Yg^*/g \mid L(t,j), C^L_{t,j}) \\
&= \int_{c_{t,j}} \int_{L(t,j)} \int_{\bar{A}(t-1),\bar{L}(t-1),L(t,1:j-1)} E_{Q,g^*}(Y \mid L(t,j), Pa(L(t,j))) \\
&I(c^L_{t,j}(Pa(L(t,j))) = c_{t,j}) \prod_{l=0}^{t-1} g^*_l \prod_{l\in\mathcal{S}(t,j)} Q_l Q_{0,t,j}(L(t,j) \mid c_{t,j}) \\
&= \int_{L(t,j)} \int_{\bar{A}(t-1),\bar{L}(t-1),L(t,1:j-1)} E_{Q,g^*}(Y \mid L(t,j), Pa(L(t,j))) \\
&\prod_{l=0}^{t-1} g^*_l \prod_{l\in\mathcal{S}(t,j)} Q_l Q_{0,t,j}(L(t,j) \mid c^L_{t,j}(Pa(L(t,j))) \\
&= E_{Q_{-(t,j)},Q_{0,t,j},g^*}Y.
\end{aligned}$$

Here we used Fubini's theorem to first integrate out $c_{t,j}$, noting that the indicator will only be non-zero at $c_{t,j} = c^L_{t,j}(\bar{L}(t-1), \bar{A}(t-1), L(t,1:j-1))$. As a consequence of the disappearing of the indicator, we end up with a pure marginal expectation w.r.t. $P_{Q_{-(t,j)},Q_{0,t,j},g^*}Y$. Similarly, we obtain $P_0E_{Q,g}(Yg^*/g \mid C^L_{t,j}) = E_{Q,g^*}Y$. So we obtain at a Q,g for which $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$ for all (t,j),

$$P_0D^*(Q,g) = \sum_{t=0}^{\tau+1} \sum_{j=1}^{n_t} E_{Q_{-(t,j)},Q_{0,t,j}-Q_{t,j},g^*}Y.$$

As shown in the previous theorem, this reduces to $\Psi(Q_0) - \Psi(Q)$ plus second order terms $R(Q,Q_0)$ The other statements are analogue to the proof of the previous theorem. This completes the proof of the theorem. □

6.2 Efficient influence curve and its robustness for our statistical model

We now present the efficient influence curve for our actual statistical model.

Theorem 3 *Consider the longitudinal data structure* $O = (L(0), A(0), \ldots, L(\tau), A(\tau), Y = L(\tau+1)) \sim P$, *and let* $L(t) = (L(t,j) : j = 1,\ldots,N)$, $t = 0,\ldots,\tau+1$.

Let $g = \prod_{t=0}^{\tau} g_t$, $g_t = P(A(t) \mid \bar{L}(t), \bar{A}(t-1))$, $Q = \prod_{t=0}^{\tau+1} \prod_{j=1}^{N} Q_{t,j}$, $Q_{t,j} = P(L(t,j) \mid L(t,1), \ldots, L(t,j-1), \bar{L}(t-1), \bar{A}(t-1))$, so that $P = Qg$. Consider the statistical model $\mathcal{M}$ that assumes a model on $g_t(A(t) \mid \bar{A}(t-1), \bar{L}(t)))$, and

$$Q_{t,j}(L(t,j) \mid L(t:1:j-1), \bar{L}(t-1), \bar{A}(t-1)) = Q_t(L(t,j) \mid c^L_{t,j})$$

for a common (in j) Q_t, and some specified functions $C^L_{t,j} = c^L_{t,j}(\bar{A}(t-1), \bar{L}(t-1))$ (with common range), respectively, $j = 1, \ldots, n$, $t = 1, \ldots, \tau+1$. Note that this corresponds with assuming that the components of $(L(t,j) : j = 1, \ldots, n)$ are conditionally independent given $\bar{L}(t-1), \bar{A}(t-1)$, with conditional distributions described by a common Q_t. In addition, it is assumed that $c^L_{t,j}$ be a d-dimensional function of $L(1:t-1,j) = (L(l,j) : l = 1, \ldots, t-1)$ and $\bar{L}(t-1)/\{L(1:t-1,j)$ for $j = 1, \ldots, N$ with d not depending on N. For notational convenience, let c^L_0 be a constant so that conditioning on this can be ignored. Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be defined as $E_{Q,g^*}Y$, where $P_{Q,g^*} = Qg^*$ is the G-computation formula for the distribution of O under stochastic intervention g^*, and $Y = 1/N \sum_{j=1}^{N} Y_j$.

Define

$$h_{t,j,Q,g}(c_{t,j}) = P_{Q,g}(C^L_{t,j} = c_{t,j}) \text{ and } h_{t,j,Q,g^*}(c_{t,j}) = P_{Q,g^*}(C^L_{t,j} = c_{t,j}).$$

Sometimes, we use short-hand notation $h_{t,j}$ for $h_{t,j,Q,g}$ and $h^*_{t,j}$ for h_{t,j,Q,g^*}. If $t = 0$, we define $h_{t=0,j,Q,g} = h_{t=0,j,Q,g^*} = 1$. Define, for each $t = 1, \ldots, \tau+1$,

$$\bar{D}_t(l(t), c(t)) =$$
$$\sum_{j=1}^{N} \left\{ E_{Q,g}(Yg^*/g \mid L(t,j) = l(t), c^L_{t,j} = c(t)) - E_{Q,g}(Yg^*/g \mid c^L_{t,j} = c(t)) \right\} \frac{h_{t,j}}{\bar{h}_t}(c(t)),$$

where $\bar{h}_t = \sum_j h_{t,j}$, $j = 1, \ldots, N$. If one assumes a common (in j) marginal distribution for $L(0,j)$, then $D^*_{t=0}(Q,g) = \sum_{j=1}^{N} \bar{D}_0(L(0,j))$, using same formula for $\bar{D}_0$ as above, with $\frac{h_{t,j}}{\bar{h}_t}(c(t)) = 1/N$. If we only assume independence of $L(0,j)$ across j, then we define

$$D^*_0(Q,g) = \sum_{j=1}^{N} \{ E_{Q,g}((Yg^*/g \mid L(0,j)) - E_{Q,g}(Yg^*/g) \}.$$

The efficient influence curve $D^*(Q,g)$ of $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ at $P_{Q,g}$ is given by

$$D^*(Q,g) = D^*_0(Q,g) + \sum_{t=1}^{\tau+1} \sum_{j=1}^{N} \bar{D}_t(L(t,j), C^L_{t,j}).$$

We have for $t = 1, \ldots, \tau+1$,

$$E_{Q,g}(Yg^*/g \mid L(t,j), C^L_{t,j}) - E_{Q,g}(Yg^*/g \mid C^L_{t,j}) =$$
$$\frac{h_{t,j,Q,g^*}}{h_{t,j,Q,g}}(C^L_{t,j}) \left\{ E_{Q,g^*}(Y \mid L(t,j), C^L_{t,j}) - E_{Q,g^*}(Y \mid C^L_{t,j}) \right\}.$$

Thus, for $t = 1, \ldots, \tau + 1$,

$$\bar{D}_t(l(t), c(t)) = \\ \sum_{j=1}^{N} \frac{h^*_{t,j}}{\bar{h}_t}(c(t)) \left\{ E_{Q,g^*}(Y \mid L(t,j) = l(t), C^L_{t,j} = c(t)) - E_{Q,g^*}(Y \mid C^L_{t,j} = c(t)) \right\},$$

so that we can represent the efficient influence curve also as:

$$D^*(Q,g) = D^*_0(Q,g) + \sum_{t=1}^{\tau+1} \sum_{j=1}^{N} \sum_{m=1}^{N} \frac{h^*_{t,m}}{\bar{h}_t}(c^L_{t,j}) \\ \left\{ E_{Q,g^*}(Y \mid L(t,m) = L(t,j), C^L_{t,m} = c^L_{t,j}) - E_{Q,g^*}(Y \mid C^L_{t,m} = c^L_{t,j}) \right\}.$$

In particular, if $E_{Q,g^*}(Y \mid L(t,j) = l(t), C^L_{t,j} = c(t)) - E_{Q,g^*}(Y \mid C^L_{t,j} = c(t))$ *is a function of* $l(t), c(t)$ *that is constant in* j*, then*

$$\bar{D}_t(l(t), c(t)) \\ = \frac{\sum_j h_{t,j,Q,g^*}}{\sum_j h_{t,j,Q,g}}(c(t)) \left\{ E_{Q,g^*}(Y \mid L(t,j) = l(t), C^L_{t,j} = c(t)) - E_{Q,g^*}(Y \mid C^L_{t,j} = c(t)) \right\} \\ = \frac{\bar{h}_{t,Q,g^*}}{\bar{h}_{t,Q,g}}(c(t)) \left\{ E_{Q,g^*}(Y \mid L(t,j) = l(t), C^L_{t,j} = c(t)) - E_{Q,g^*}(Y \mid C^L_{t,j} = c(t)) \right\},$$

where $\bar{h}_{t,Q,g} \equiv \sum_{j=1}^{N} h_{t,j,Q,g}$.

Robustness of efficient influence curve: *Let* $P_0 = Q_0 g_0$. *If* $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$ *for* $j = 1, \ldots, N$, *and* $Q_t = Q_{0,t}$, *then* $P_0 D^*_{t,j}(Q,g) = 0$. *Specifically, if* $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$ *for all* $j = 1, \ldots, N$, *then* $P_0 D^*_{t,j}(Q,g) = E_{Q_{-t},Q_{0,t},g^*}Y - E_{Q,g^*}Y = 0$.

If $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$ *for all* t, j, *then*

$$P_0 D^*(Q,g) = E_{Q_{0,0},Q,g^*}Y - E_{Q,g^*}Y + \sum_{t=1}^{\tau+1} \sum_{j=1}^{N} \{E_{Q_{-t},Q_{0,t},g^*}Y - E_{Q,g^*}Y\}.$$

Define

$$R(Q, Q_0) \equiv \sum_{t=0}^{\tau+1} \sum_{j=1}^{N} \{E_{Q_{0:t-1} - Q_{0,0:t-1}, Q_{0,t} - Q_t, Q_{t+1:\tau+1}, g^*}Y.$$

Note that $R(Q, Q_0)$ *is a second order term involving square differences between* Q *and* Q_0. *This implies that, if* $h_{t,j,Q,g} = h_{t,j,Q_0,g_0}$ *for all* t, j, *then*

$$P_0 D^*(Q,g) = \Psi(Q_0) - \Psi(Q) + R(Q_0, Q).$$

Define

$$R((h,h_0),(Q,Q_0)) \equiv \sum_t \sum_m \int_{l(t),c(t)} h^*_{t,m} Q_{0,t} \frac{(\bar{h}_{t,0} - \bar{h}_t)}{\bar{h}_t} \\ \left\{ E_{Q,g^*}(Y \mid L(t,m) = l(t), C^L_{t,m} = c(t)) - E_{Q,g^*}(Y \mid C^L_{t,m} = c(t) \right\}$$

Since the conditional expectation w.r.t. $Q_{0,t}$ of the difference of the two conditional expectations equals zero if $Q_t = Q_{0,t}$, it follows that the t-specific term on the right-hand side involves a difference $Q_{0,t} - Q_t$, and thereby a product $(Q_{0,t} - Q_t)(\bar{h}_t - \bar{h}_{0,t})$. In general, we have

$$P_0 D^*(Q, g) = \Psi(Q_0) - \Psi(Q) + R((h, h_0), (Q, Q_0)) + R(Q, Q_0).$$

In particular, for all g, $P_0 D^*(Q_0, g) = 0$.

Proof. The proof is analogue to the proof of the previous theorem and uses the following additional lemma. □

Lemma 1 *Let $T_{Q_k} = \{\sum_j S(L(k,j) \mid c^L_{k,j}) : \int S(l \mid c) Q_k(l \mid c) = 0\}$ be the tangent space of Q_k. Recall that the marginal density of $C^L_{k,j}$ is given by h_j, $j = 1, \ldots, N$. Let D be an element of $L^2_0(P)$. The projection of D onto T_{Q_k} is given by*

$$\Pi(D \mid T_{Q_k}) = \sum_{j=1}^{N} \bar{D}_k(L(k,j) \mid C^L_{k,j}),$$

where

$$\bar{D}_k(l \mid c) = \frac{1}{\sum_j h_j(c)} \sum_j \{E(D \mid L(k,j) = l, C^L_{k,j} = c) - E(D \mid C^L_{k,j} = c)\} h_j(c).$$

Proof Lemma: The proof of this lemma consists of two steps. Firstly, one notes that the conjectured projection is indeed an element of the desired tangent space. Secondly, one shows that D minus the conjectured projection is orthogonal to any element in T_{Q_k}. Define $D_j(L(k,j) \mid c^L_{k,j}) = E(D \mid L(k,j), c^L_{k,j}) - E(D \mid c^L_{k,j})$.

We note that T_{Q_k} is embedded in the larger sub-Hilbert space $\{\sum_j S_j(L(k,j) \mid C^L_{k,j}) : \int S_j(l \mid c) Q_k(l \mid c) = 0\}$. The projection of D onto this larger space is given by $\sum_j D_j$. Thus, the projection of D onto T_{Q_k} is given by $\Pi(\sum_j D_j \mid T_{Q_k})$.

Let the conjectured projection be $\sum_j \bar{D}_k(L(k,j) \mid c^L_{k,j})$. In the remainder of the proof we suppress the index k which plays no role. The orthogonality requirement states that for each f

$$E\left\{\sum_j D_j - \sum_j \bar{D}\right\} \sum_j f = 0.$$

Thus, for all functions f we have

$$\textstyle\int_{l,c} \{\sum_j D_j(l,c) - \sum_j \bar{D}(l,c)\} \sum_j f(l,c) Q(l \mid c) h_j(c) = 0.$$

This can be rewritten as:

$$\int_{l,c}\left\{\sum_j D_j(l,c)h_j(c)\right\} f(l,c)Q(l,c) = \int_{l,c}\bar{D}(l,c)\sum_j h_j(c)f(l,c)Q(l \mid c).$$

Since this equality needs to hold for all functions f, it follows that

$$\sum_j D_j(l,c)h_j(c) = \bar{D}(l,c)\sum_j h_j(c),$$

and thus

$$\bar{D}(l,c) = \sum_j D_j(l,c)h_j(c)/\sum_j h_j(c).$$

This completes the proof of the lemma. □

The second order terms We note that the second order terms will only be second order if $E(Y_j \mid L(t,m), C^L_{t,m}) - E(Y_j \mid C^L_{t,m}) = 0$ for most m (or very small for most m), so that substituting $Y = 1/N\sum_j Y_j$ in the second order expressions indeed results in a stable average of second order differences.

6.3 Generalization of case that the individuals are independent

We demonstrate that the last theorem generalizes the efficient influence curve for i.i.d. data in which case the statistical graph does not make any more conditional independence assumptions than implied by the time-ordering, beyond the independence of all units.

Corollary 1 *Consider longitudinal data structure* $O = (L(0), A(0), \ldots, L(\tau), A(\tau), L(\tau+1)) \sim P$, *and let* $L(t) = (L(t,j) : j = 1,\ldots,N)$, $A(t) = (A(t,j) : j = 1,\ldots,N)$, $k = 0,\ldots,\tau+1$. *Let* $Y = \frac{1}{N}\sum_j L(\tau+1,j)$. *Let* $g = \prod_{t=0}^{\tau} g_t$, $g_t = P(A(t) \mid \bar{L}(t), \bar{A}(t-1))$, $Q = \prod_{t=0}^{\tau+1}\prod_{j=1}^{N} Q_{t,j}$, $Q_{t,j} = P(L(t,j) \mid L(t,1),\ldots,L(t,j-1), \bar{L}(t-1), \bar{A}(t-1))$, *so that* $P = Qg$. *Consider the statistical model* $\mathcal{M}$ *that assumes*

$$g_t(A(t) \mid \bar{A}(t-1), \bar{L}(t-1)) = \prod_j g_t(A(t,j) \mid \bar{A}_j(t-1), \bar{L}_j(t-1)),$$

and

$$Q_{t,j}(L(t,j) \mid L(t:1:j-1), \bar{L}(t-1), \bar{A}(t-1)) = Q_t(L(t,j) \mid \bar{L}_j(t-1), \bar{A}_j(t-1))$$

for a common (in j) Q_t, $j = 1, \ldots, n$, $t = 1, \ldots, \tau + 1$. Thus, we now have $C^L_{t,j} = (\bar{A}_j(t-1), \bar{L}_j(t-1))$, $C^A_{t,j} = (\bar{A}_j(t-1), \bar{L}_j(t))$, only includes the history of the j-th individual. The above model corresponds with assuming that $O = (O_1, \ldots, O_N)$, $O_j = (L_j(0), A_j(0), \ldots, L_j(\tau), A_j(\tau), Y_j)$, and O_j are independent and identically distributed with common distribution $\prod_t g_t \prod_t Q_t$.

For notational convenience, let c^L_0 be a constant so that conditioning on this can be ignored. Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be defined as $E_{Q,g^*}Y$, where $P_{Q,g^*} = Qg^*$ is the G-computation formula for the distribution of O under stochastic intervention g^*, and $Y = 1/N \sum_{j=1}^N Y_j$.

Define

$$h_{t,j,Q,g}(c_{t,j}) = P_{Q,g}(C^L_{t,j} = c_{t,j}) \text{ and } h_{t,j,Q,g^*}(c_{t,j}) = P_{Q,g^*}(C^L_{t,j} = c_{t,j}).$$

Define, for each $t = 0, \ldots, \tau + 1$,

$$\bar{D}_t(l(t), c(t)) = \\ \sum_{j=1}^N \left\{ E_{Q,g}(Yg^*/g \mid L(t,j) = l(t), C^L_{t,j} = c(t)) - E_{Q,g}(Yg^*/g \mid C^L_{t,j} = c(t)) \right\} \frac{h_{t,j}}{\bar{h}_t}(c(t)),$$

where $\bar{h}_t = \sum_j h_{t,j}$, and $h_{t,j}(c) = P(C^L_{t,j} = c)$ is the marginal density of $C^L_{t,j}$, $j = 1, \ldots, N$. The previous theorem establishes that the efficient influence curve $D^*(Q,g)$ of $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ at $P_{Q,g}$ is given by

$$D^*(Q,g) = \sum_{t=0}^{\tau+1} \sum_{j=1}^N \bar{D}_t(L(t,j), C^L_{t,j}),$$

where

$$\begin{aligned} &\bar{D}_t(l(t), c(t)) \\ &= \frac{\sum_j h_{t,j,Q,g^*}}{\sum_j h_{t,j,Q,g}}(c(t)) \left\{ E_{Q,g^*}(Y \mid L(t,j) = l(t), C^L_{t,j} = c(t)) - E_{Q,g^*}(Y \mid C^L_{t,j} = c(t)) \right\} \\ &= \frac{\bar{h}_{t,Q,g^*}}{\bar{h}_{t,Q,g}}(c(t)) \left\{ E_{Q,g^*}(Y \mid L(t,j) = l(t), C^L_{t,j} = c(t)) - E_{Q,g^*}(Y \mid C^L_{t,j} = c(t)) \right\}, \end{aligned}$$

and $\bar{h}_{t,Q,g} \equiv \sum_{j=1}^N h_{t,j,Q,g}$.

We have $h_{t,j,Q,g}(\bar{a}_j(t-1), \bar{l}_j(t-1)) = Qg(\bar{a}_j(t-1), \bar{l}_j(t-1))$, so that $h_{t,j,Q,g^*}/h_{t,j,Q,g}(c^L_{t,j}) = \prod_{s \leq t} g^*_s(a_j(s) \mid \bar{a}_j(s-1), \bar{l}_j(s)) / \prod_{s \leq t} g_s(a_j(s) \mid \bar{a}_j(s-1), \bar{l}_j(s))$. In addition,

$$\begin{aligned} &E_{Q,g^*}(Y \mid L(t,j), C^L_{t,j}) - E_{Q,g^*}(Y \mid C^L_{t,j}) \\ &= \frac{1}{N} \left\{ E_{Q,g^*}(Y(j) \mid L(t,j), \bar{L}_j(t-1), \bar{A}_j(t-1)) - E_{Q,g^*}(Y(j) \mid \bar{L}_j(t-1), \bar{A}_j(t-1)) \right\} \end{aligned}$$

is constant in j. Thus, the efficient influence curve reduces to

$$D^*(Q,g)(O) = \frac{1}{N} \sum_{t=0}^{\tau+1} \sum_{j=1}^N \bar{D}_t(L(t,j), C^L_{t,j}),$$

where

$$\bar{D}_t(L(t,j),C^L_{t,j}) = \frac{g^*}{g}(\bar{A}_j(t-1),\bar{L}_j(t-1))\{E_{Q,g^*}(Y(j)\mid L(t,j),C^L_{t,j}) - E_{Q,g^*}(Y(j)\mid C^L_{t,j})\}.$$

7 Characterizing the optimal asymptotic variance of the MLE in terms of efficient influence curve.

Due to our sequential conditional independence assumption, the log-likelihood can be represented as a sum over time-points t and units i, and for each t, the sum over i consists of independent random variables, conditional on the past. As a consequence, under regularity conditions, one can show that the log-likelihood is asymptotically normally distributed. Therefore, we conjecture that we can establish so called local asymptotic normality of our statistical model, which involves establishing asymptotic normality of log-likelihood under sampling from fluctuations/submodels $P_{\epsilon=1/\sqrt{N}} \subset \mathcal{M}$ of a fixed data distribution P across all possible fluctuations. As shown in van der Vaart (1998), for models satisfying the local asymptotic normality condition, the normal limit distribution of an MLE is an optimal limit distribution based on the convolution theorem. In this section we informally demonstrate the importance of the efficient influence curve as the random variable whose variance characterizes the normal limit distribution of an MLE of the target parameter for our semiparametric network model for $N \rightarrow \infty$, and thereby characterizes the normal limit distribution of optimal estimators. As part of this we use a template for establishing the normal limit distribution of the MLE, which can be equally well applied to the TMLE in the next section using the true g_0.

Even though it is well known that a regular estimator based on sample of n i.i.d observations is efficient if and only if it is asymptotically linear with influence curve equal to the efficient influence curve, here we are not interested in asymptotics when we observe n of our data structures that are indexed by this parameter N (like observing n networks of size N individuals), but we are interested in the asymptotics in N based on a single draw of O. Therefore, we think it is important to point out that the asymptotic behavior of the MLE based on such a single O when $N \rightarrow \infty$, showing that the asymptotic variance is still characterized by the efficient influence curve. Our lesson is that our goal should still be to construct an estimator that is asymptotically normally distributed with variance equal to the variance of the efficient influence curve, appropriately normalized.

Specifically, we show that, under appropriate regularity conditions, the asymptotic variance of a standardized MLE $\sqrt{N}(\psi_N - \psi_0)$ of the target parameter (assuming it is well defined asymptotically) equals the limit in N of $NP_0\{D^*(Q_0, g_0)\}^2$, where $P_0\{D^*(Q_0, g_0)\}^2$ is the variance of the efficient influence curve $D^*(Q_0, g_0)$. The formal analysis of an MLE is almost as hard as a formal analysis of the TMLE since it requires understanding an empirical process $Z_N(Q)$ specified below uniformly in Q, which is challenging due to the fact that, contrary to $Z_N(Q_0)$, at misspecified Q, the time-specific components of $Z_N(Q)$ cannot be represented as sums of independent random variables, conditional on the history at that time. Since the TMLE is tailored to deal with the curse of dimensionality (and MLE is a special case of TMLE by defining the initial estimator for the TMLE as the MLE, assuming this MLE is well defined estimator), we consider the analysis of a TMLE more important. Such a formal analysis is presented for the point-treatment $K = 0$-case in a later section and much can be learned from that analysis for the purpose of analyzing the TMLE or MLE for general K. Nonetheless, the template below can be used to establish the asymptotic normality for both the MLE and the TMLE.

Let Q_N be an MLE, assuming it is well defined (i.e., all covariates are discrete and of low dimension relative to N). We wish to analyze the plug-in MLE $\Psi(Q_N)$ of ψ_0. We represent the efficient influence curve as $D^*(Q, g) = D^*(Q, h(Q, g), \Psi(Q))$. Recall $P_0 D^*(Q_N, h(g_0, Q_N), \Psi(Q_N)) = (\psi_0 - \Psi(Q_N)) + R_N$, where R_N is a second order term defined as sum of two terms $R_N(Q_N, Q_0)$ and $R_N((h_N, h_0), (Q_N, Q_0))$. The first involves square differences of Q_N, Q_0, while the second involves the product of differences $h(g_0, Q_N) - h(g_0, Q_0)$ and $Q_N - Q_0$. We will assume that a separate analysis establishes that $R_N = o_P(1/\sqrt{N})$. Since Q_N is an MLE, and $D^*(Q_N, h(g_0, Q_N), \Psi(Q_N))$ is a score at P_{g_0,Q_N}, we have that is solves the efficient influence curve equation, $D^*(Q_N, h(g_0, Q_N), \Psi(Q_N)) = 0$. We also have $P_0 D^*(Q_0, h, \psi_0) = 0$ for any h. This allows us to establish a first order expansion of the standardized MLE:

$$\begin{aligned}&(\Psi(Q_N) - \psi_0) = -P_0 D^*(Q_N, h(g_0, Q_N), \Psi(Q_N)) + R_N\\&= D^*(Q_N, h(g_0, Q_N), \Psi(Q_N)) - P_0 D^*(Q_N, h(g_0, Q_N), \Psi(Q_N)) + R_N.\end{aligned}$$

Thus under the assumption that $R_N = o_P(1/\sqrt{N})$, it follows that the asymptotic distribution of $\sqrt{N}(\Psi(Q_N) - \Psi(Q_0))$ equals the limit distribution of

$$Z_N(Q_N) = \sqrt{N}(D^*(Q_N, h(g_0, Q_N), \Psi(Q_N)) - P_0 D^*(Q_N, h(g_0, Q_N), \Psi(Q_N))).$$

A non-trivial analysis as carried out for the case $\tau = 1$, and using appropriate conditions, can be used to establish that $Z_N(Q_N) - Z_N(Q_0) = o_P(1)$, so that $Z_N(Q_N)$ behaves as $Z_N(Q_0)=\frac{1}{\sqrt{N}}(D^*(Q_0,h_0,\psi_0)-P_0D^*(Q_0,h_0,\psi_0))$. It remains to investigate weak convergence of $Z_N(Q_0)$ as N converges to infinity.

For notational convenience, let $D^*(Q_0)=D^*(Q_0,h_0,\psi_0)$. We have the following representation from Theorem 3:

$$D^*(Q_0,g_0)=\tfrac{1}{N}\textstyle\sum_t\sum_j\sum_m \frac{h^*_{t,m}}{\bar{h}_t}(C^L_{t,j})\sum_l D_{l,t,m}(L(t,j),C^L_{t,j}),$$

where

$$D_{l,t,m}=E(Y(l)\mid L(t,m)=L(t,j),C^L_{t,m}=C^L_{t,j})-E(Y(l)\mid C^L_{t,m}=C^L_{t,j}).$$

We can represent this as

$$D^*(Q_0,g_0)=\frac{1}{N}\sum_t\sum_j D^*_t(Q_0,g_0)(L_{t,j},C^L_{t,j}),$$

where we defined

$$D^*_t(Q_0,g_0)(L(t,j),C^L_{t,j})=\left\{\sum_m\frac{h^*_{t,m}}{\bar{h}_t}(C^L_{t,j})\sum_l D_{l,t,m}(L(t,j),C^L_{t,j})\right\}$$

Note that $D^*_t(Q_0,g_0)$ has conditional mean zero, given $C^L_{t,j}$. In order to claim that $D^*_t(Q_0,g_0)$ has finite variance one needs that the summation over l reduces essentially to a finite sum due to $L(t,m)$ being conditionally independent of $Y(l)$, given $C^L_{t,m}$, for most m.

This yields the following representation:

$$Z_N(Q_0)=\frac{1}{\sqrt{N}}\sum_{t,i}D^*_t(Q_0)(L(t,i),C^L_{t,i})),$$

where D^*_t is a function of $L(t,i)$ and $C^L_{t,i}$ with conditional mean zero, given $C^L_{t,i}$. Due to factorization of the likelihood as $Q=\prod_{t,i}Q_{L_i(t)\mid C^L_{t,i}}$, it follows that the variance of $Z_N(Q_0)$ is given by

$$\mathrm{VAR}Z_N(Q_0)=\frac{1}{N}\sum_{t,i}P_0\{D^*_t(Q_0)(L(t,i),C^L_{t,i})\}^2.$$

We have

$$P_0\{D^*_t(Q_0)(L(t,i),C^L_{t,i})\}^2=\int_{l(t),c(t)}D^*_t(Q_0)(l(t),c(t))^2Q_{0,t}(l(t)\mid c(t))h_{0,t,i}(c(t)).$$

Thus, the asymptotic variance of $Z_N(Q_0)$ is given by limit of

$$\sigma_0^2 \equiv \lim_{N\to\infty} \sum_t \int_{l(t),c(t)} \{D_t^*(Q_0)(l(t),c(t))\}^2 Q_{0,t}(l(t) \mid c(t)) \frac{\bar{h}_{0,t}}{N}(c(t)),$$

where $\bar{h}_t = \sum_i h_{t,i}$. Note that we also have that $\sigma_0^2 = \lim_{N\to\infty} NP_0D^*(Q_0)^2$ equals N times the variance of the efficient influence curve $D^*(Q_0)$ for the target parameter Ψ at P_0.

In addition, we note that $Z_N(Q_0) = \sum_t Z_{N,t}(Q_0)$, where $Z_{N,t}(Q_0) = 1/\sqrt{N}\sum_i D_{t,i}^*(Q_0)$, with $D_{t,i}^* = D_t^*(L(t,i), C_{t,i}^L)$, is a sum of independent random variables $L(t,i)$, conditional on $\bar{A}(t-1), \bar{L}(t-1)$. By using CLT theorems, we can therefore establish that for each $t = 0,\ldots,\tau+1$, $Z_{N,t}(Q_0)$ converges weakly to a normal distribution $Z_t(Q_0)$. These t-specific limit random variables $Z_t(Q_0)$ are pairwise independent, so that the sum across t converges to a normal distribution with variance equal to the sum of the t-specific variances, and thus σ_0^2 as defined above. To conclude, under appropriate regularity conditions, we will have that $Z_N(Q_0)$ converges weakly to $N(0,\sigma_0^2)$

This demonstrates that the efficient influence curve characterizes the limit distribution of the maximum likelihood estimator, and thus indeed characterizes asymptotically optimal mean zero normal limit distribution identified by the asymptotic variance of $Z_N(Q_0)$.

8 TMLE

Consider the likelihood $L(Q) = \prod_{t,j} Q_{t,j}$ and recall the representation of the efficient influence curve $D^*(Q,g) = \sum_t \frac{1}{N}\sum_j D_{t,j}^*(Q,g)$, where $D_{t,j}^*(Q,g) = D_t^*(Q,g)(L(t,j), C_{t,j}^L)$. Let Q_N and g_N be an initial estimator. Let $Q_{t,j,N}(\epsilon)$ be a fluctuation of $Q_{t,j,N}$ with score at $\epsilon = 0$ equal to $D_{t,j}^*(Q_N, g_N)$ such as $Q_{t,j,N}(\epsilon) = C(\epsilon)\exp(\epsilon D_{t,j}^*(Q_N,g_N))Q_{t,j,N}$. Recall that $D_{t,j}^*$ is a common function of $(L(t,j), C_{t,j}^L)$ so that this submodel is indeed contained in our model. Let $\epsilon_{t,N}$ be the maximum likelihood estimator, and let $Q_{t,j}(\epsilon_{t,j,N})$, $j = 1,\ldots,N$, be the update $Q_{t,j}^1$, $j = 1,\ldots,N$. Carry this update out for each t, which provides a mapping from an initial Q_N^0 into an update Q_N^1. Iterate this updating process till convergence, denote the final updates with $(Q_{t,j,N}^* : t,j)$, and let Q_N^* be the corresponding TMLE of Q_0. The TMLE of ψ_0 is now the plug-in estimator $\Psi(Q_N^*)$.

By construction, the TMLE solves the efficient influence curve equation $D^*(Q_N^*, g_N)(O) = 0$. In the previous section we presented a template for the analysis of MLE that can be equally well applied to this TMLE for the case that $g_N = g_0$.

8.1 External estimation of h_0

The efficient influence curve depends on the data generating distribution P through $Q(P), h(P)$ only. Given estimators of Q_N, g_N, one obtains a plug-in estimator $h(Q_N, g_N)$ of $h_0 = h(P_0)$. The second order term for the standardized TMLE, and thereby the bias of the TMLE, involves square differences of $Q_N - Q_0$, and a product of differences $h_N - h_0$ and $Q_N - Q_0$. This suggests that the only goal in estimation of g_0 is to construct a good estimator of h_0. The important advantage of this plug-in estimator $h(Q_N, g_N)$ of h_0 is that it fully utilizes the knowledge coded by the statistical model.

Nonetheless, it may be of interest to construct a direct estimator of h_0 separate from an estimator Q_N of Q_0. For that purpose we note that $\bar{h}_{t,0}$ is a density of $C^L_{t,j}$. We can estimate $\bar{h}_{t,0}$ by using a density estimator treating $C^L_{t,j}$, $j = 1, \ldots, N$, as i.i.d. This corresponds with using as loss-function for $\bar{h}_{t,0}$

$$L(\bar{h}_t)(O) = -\sum_j \log \bar{h}_t(C^L_{t,j}).$$

Ad hoc refinements based on recognizing that $C^L_{t,j}$ are dependent across j can be based on utilizing the measured connectivity between the N individuals. It remains to be investigated how these two methods of estimation of $\bar{h}_0$, plug-in versus direct estimation, compare w.r.t. behavior of the resulting estimator of ψ_0.

8.2 TMLE relying on external estimator of h_0.

Consider the likelihood $L(Q) = \prod_{t,j} Q_{t,j}$ and recall the representation of the efficient influence curve $D^*(Q,h) = \sum_t \frac{1}{N} \sum_j D^*_{t,j}(Q,h)$. We note that even if h is not compatible with Q, $D^*_{t,j}(Q,h)$ is still a score of $Q_{t,j}$ (i.e., it is a function of $L(t,j)$ and $C^L_{t,j}$ with conditional mean zero, given $C^L_{t,j}$). This suggests that in our fluctuations of the TMLE we could use an external estimator h_N of h_0, and not update h_N in the TMLE updating algorithms. This results in the following TMLE.

Let Q_N and h_N be an initial estimator. Let $Q_{t,j,N}(\epsilon)$ be a fluctuation of $Q_{t,j,N}$ with score at $\epsilon = 0$ equal to $D^*_{t,j}(Q_N, h_N)$ such as $Q_{t,j,N}(\epsilon) = C(\epsilon)\exp(\epsilon D^*_{t,j}(Q_N, h_N))Q_{t,j,N}$. Recall that $D^*_{t,j}$ is a common function of $L(t,j), C^L_{t,j}$ so that this submodel is indeed contained in our model. Let $\epsilon_{t,N}$ be the maximum likelihood estimator, and let $Q_{t,j}(\epsilon_{t,j,N})$, $j = 1, \ldots, N$, be the update $Q^1_{t,j}$, $j = 1, \ldots, N$. Carry this update out for each t, which provides a mapping from an initial Q^0_N into an update Q^1_N. Iterate this updating process till convergence and denote the final updates with $(Q^*_{t,j,N} : t, j)$, and let Q^*_N be the corresponding TMLE of Q_0. The TMLE of ψ_0 is now the plug-in estimator $\Psi(Q_N^*)$. By construction, the TMLE solves the efficient influence curve equation $D^*(Q_N^*, h_N)(O) = 0$. Therefore the previously presented template for analyzing the MLE and TMLE can be applied to this TMLE as well.

8.3 A simple TMLE

Consider the likelihood $L(Q) = \prod_{t,j} Q_{t,j}$. Let Q_N and h_N be an initial estimator. Suppose that under Q_N $L(t,j)$ is a deterministic function of $C_{t,j}^L$ for all $j = 1, \ldots, N$, $t = 1, \ldots, \tau + 1$. At such a Q_N we have that $D_{t,j}^* = 0$ for all $t = 1, \ldots, \tau$. Under the assumption that $L_i(0)$ are i.i.d. (note, $\bar{h}_{t=0} = N$), we also have $D_{t=0,j}^*(Q,h)(L(0,j)) = \frac{1}{N} \sum_{m=1}^N E_{Q,g^*}(Y \mid L(0,m) = L(0,j)) - \Psi(Q)$, where $Y = \frac{1}{N} \sum_{i=1}^N L(\tau + 1, i)$.

We have

$$D_{\tau+1,j}^*(Q,h) = \sum_{m=1}^N \frac{h_{\tau+1,m}^*}{\bar{h}_{\tau+1}} (c_{\tau+1,j}^L) \int_{L(0),A} \frac{1}{N} \left\{ Y_j - Y_m^*(Q)(L(0),A) \right\} \\ \frac{I(c_{\tau+1,m}^L(\bar{A}(\tau),\bar{L}(\tau)) = c_{\tau+1,j}^L)}{h_{\tau+1,m,Q,g}(c_{\tau+1,j}^L)} \prod_{l=0}^K g_l^* \prod_{l \in S(\tau+1,j)} Q_l$$

Let $Q_{\tau+1,j,N}(\epsilon)$ be a fluctuation of $Q_{\tau+1,j,N}$ with score at $\epsilon = 0$ equal to $D_{\tau+1,j}^*(Q_N, h_N)$ such as $Q_{\tau+1,j,N}(\epsilon) = C(\epsilon) \exp(\epsilon D_{\tau+1,j}^*(Q_N, h_N)) Q_{\tau+1,j,N}$. Let $\epsilon_{\tau+1,N}$ be the maximum likelihood estimator, and let $Q_{\tau+1,j,N}(\epsilon_{\tau+1,N})$ be the update $Q_{\tau+1,j,N}^1$. Carry this update out for each t, which provides a mapping from an initial $Q_{\tau+1,j,N}^0$, $j = 1, \ldots, N$, into an update $Q_{\tau+1,j,N}^1$, $j = 1, \ldots, N$. Iterate this updating process for $(Q_{\tau+1,j} : j)$ till convergence and denote the final update with $(Q_{\tau+1,j,N}^* : j)$, and let Q_N^* be the corresponding TMLE of Q_0, which thus only updates $Q_{\tau+1,j,N}^0$, $j = 1, \ldots, N$. The TMLE of ψ_0 is now the plug-in estimator $\Psi(Q_N^*)$. The TMLE solves $D^*(Q_N^*, h_N)(O) = 0$.

9 The TMLE of Causal Effect of Single Time Point Intervention

We will present the TMLE for the point-treatment intervention case (i.e., $\tau = 0$). This case is of great interest itself, extends estimation of a causal effect of a single time point intervention to dependent data of the form studied in this article, and thereby covers important applications. In the next section we will formally analyze this TMLE. The tools of the proof will be generalizable to the general τ case. In addition, the single time point case allows for a TMLE that is actually double robust in the sense that it remains consistent if either Q_0 or h_0 is consistently estimated.

9.1 NPSEM

The NPSEM states:

$$\begin{aligned}
&W_i = L_i(0) = f_{W_i}(U_{W_i})\\
&A_i = A_i(0) = f_A(c_i^A(W), U_{A_i})\\
&Y_i = L_i(1) = f_Y(c_i^Y(W, A), U_{Y_i})\\
&i = 1, \ldots, N,
\end{aligned}$$

where $c_i^A(W)$ is determined by $W = (W_1, \ldots, W_N)$, and $c_i^Y(W, A)$ is determined by W, A with $A = (A_1, \ldots, A_N)$. The "friends" F_i of subject i may be included in W_i: $F_i \subset W_i$. Sometimes, we denote the baseline covariates explicitly with the notation (F_i, W_i) to stress the inclusion of F_i. The function $c_i^A(W)$ includes W_i, beyond summary measures of $(W_j : j \in F_i)$, and might be defined as $(W_i, (W_j : j \in F_i))$, assuming that $\mid F_i \mid \leq K < \infty$ for some fixed K, so that $c_i^A(W)$ can indeed be defined as a fixed multivariate dimensional function not depending on N. Similarly, the function $c_i^Y(W, A)$ includes (W_i, A_i), and might be defined as $(W_i, A_i, ((W_j, A_j) : j \in F_i))$. The above structural equation model assumes that A_i and Y_i are the same function of this dimension reduction $(W_i, (W_j : j \in F_i))$ and $(W_i, A_i, (W_j, A_j : j \in F_i))$ for each i, so that two units with the same number of friends who have the same individual covariate and treatment values, and also have the same values for the covariates and treatments of their friends, will be subjected to the same conditional distribution for drawing their treatment and outcome. In our asymptotics theorem in the next section we treat F_i, $i = 1, \ldots, N$, as fixed, so that also the probability distribution of O and the target parameter ψ_0 are indexed by the fixed values of $(F_i : i = 1, \ldots, N)$.

In addition, we assume that $U_1, \ldots, U_N$ are i.i.d: note that (since f_{W_i} is allowed to be different for each i) this corresponds with assuming that $W_1, \ldots, W_N$ are independent, but not necessarily identically distributed. This independence assumption on the U_i's implies that 1) $W_1, \ldots, W_N$ are independent, 2) conditional on $W = (W_1, \ldots, W_N)$, $A_1, \ldots, A_N$ are independent, and, 3) conditional on (W, A), $Y_1, \ldots, Y_N$ are independent. Thus, all the dependence between units is not due to the dependence of the errors, but only due to the interdependence between units as described by the structural equations that allow that an individual's treatment and outcome are a function of the baseline covariates and treatments of its friends.

Causal quantity: Let g^* be a conditional distribution of A, given W. Our goal is to estimate the mean of the counterfactual outcome of $Y^c = 1/N \sum_{i=1}^N Y_i$ under the stochastic intervention g^*. Let $Y_{g^*} = (Y_{g^*,i} : i = 1, \ldots, N)$ be the counterfactual indexed by a stochastic intervention g^* on A, and $Y^c_{g^*} = 1/N \sum_{i=1}^N Y_{g^*,i}$. The causal quantity of interest is defined as $\Psi^F(P_{U,W,A,Y}) = EY^c_{g^*}$, which is a parameter of the distribution of (U, W, A, Y).

Identifiability: We observe $O = (O_1, \ldots, O_N)$, where $O_i = (W_i, A_i, Y_i)$. Due to the above assumptions, the probability distribution of O is given by:

$$P(O) = \prod_{i=1}^N P_{W_i}(W_i) P_{A|C^A}(A_i \mid C^A_i) P_{Y|C^Y}(Y_i \mid C^Y_i), \qquad (3)$$

where $P_{A|c^A}(\cdot \mid c^A)$ is a common (in i) density for A for each c^A, and $P_{Y|C^Y}(\cdot \mid c^Y)$ is a common density for Y for each c^Y. If we assume the randomization assumption stating that $A = (A_1, \ldots, A_N)$ is independent of $U_Y = (U_{Y_i} : i = 1, \ldots, N)$, given $W = (W_1, \ldots, W_N)$, then the post-intervention probability distribution P_{g^*} of $(W, Y_{g^*}) = (W_i, Y_{i,g^*} : i = 1, \ldots, N)$ is identified by the following G-computation formula applied to the probability distribution P of O:

$$\begin{aligned} P_{g^*}(W, A^*, Y) &= \prod_{i=1}^N P_{W_i}(W_i) P_Y(Y_i \mid C_i^{Y,*}) g_i^*(A_i^* \mid C_i^{A,*}) && (4) \\ &\equiv P^{g^*}(W, A^*, Y), && (5) \end{aligned}$$

where $P_{g^*,Y}$ is defined by the conditional distribution of Y_i, given C^Y_i, which is constant in $i = 1, \ldots, N$, and where A in the parents $C^Y_i(A, W)$ is replaced by A^*. Recall our notation $C^Y_i = c^Y_i(W, A)$, $C^A_i = c^A_i(W)$. We denoted the right-hand side in this G-computation formula with P^{g^*}, which is thus always defined as a parameter of the data distribution P for a P in the statistical model implied by our causal model. In most applications one will have that the conditional distribution g^*_i of A_i, given C^A_i, is constant in i, and that under g^* $(A_1, \ldots, A_N)$ are conditionally independent, given W.

Statistical model and statistical target parameter: Let $\mathcal{M}$ be the statistical model for the data distribution P defined by (3) in which P_{W_i}, $i = 1, \ldots, N$, and the common $P_{A|C^A}$, $P_{Y|C^Y}$ are unspecified. Let the statistical target parameter mapping $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be defined as $\Psi(P) = E_{P^{g^*}} Y^{g^*,c}$. Under the stated causal model and identifiability assumptions under which $P = P_{P_{U,W,A,Y}}$, we have $\Psi(P) = \Psi^F(P_{U,W,A,Y})$, so that $\Psi(P)$ can be interpreted as the desired causal quantity. Our goal is to construct an estimator of $\psi_0 = \Psi(P_0)$ based on $O = (O_1, \ldots, O_N) \sim P_0 \in \mathcal{M}$.

Let Q_{W_i} be the marginal distribution of W_i and let Q_Y be the common conditional distribution of Y_i, given C^Y_i. The target parameter $\Psi(P)$ only depends on P through Q_{W_i}, $i = 1, \ldots, N$, and Q_Y. If we want to emphasize that $\Psi(P)$ only depends on P through $Q(P) = ((Q_{W_i} : i = 1, \ldots, N), Q_Y)$, then we will also use (and abuse) the notation $\Psi(Q)$ to indicate the mapping from Q into the desired estimand.

9.2 Efficient influence curve

Recall the general representation of the efficient influence curve for the longitudinal data structure:

$$D^*(Q,g) = \textstyle\sum_{j=1}^N \{E_{Q,g^*}(Y^c \mid L(0,j)) - E_{Q,g^*} Y^c\} \\ + \textstyle\sum_{t=1,j,m} \frac{h^*_{t,m}}{h_t}(C^L_{t,j}) \left\{E_{Q,g^*}(Y^c \mid L(t,m) = L(t,j), C^L_{t,m} = C^L_{t,j}) - E_{Q,g^*}(Y^c \mid C^L_{t,m} = C^L_{t,j})\right\}.$$

We now have $\tau = 0$, giving the following two terms:

$$D^*(Q,g) = \textstyle\sum_{j=1}^N \{E_{Q,g^*}(Y^c \mid L(0,j)) - E_{Q,g^*} Y^c\} \\ + \textstyle\sum_{j=1}^N \sum_{m=1}^N \frac{h^*_m}{h}(C^Y_j) \left\{E_{Q,g^*}(Y^c \mid Y(m) = Y(j), C^Y_m = C^Y_j) - E_{Q,g^*}(Y^c \mid C^Y_m = C^Y_j)\right\}.$$

We have

$$E^*(Y^c \mid Y_m, C^Y_m) = \tfrac{1}{N} \textstyle\sum_{j \neq m} E^*(Y_j \mid Y_m, C^Y_m) + 1/NY_m \\ = \tfrac{1}{N} \textstyle\sum_{j \neq m} E^*(E^*(Y_j \mid Y_m, W, A) \mid Y_m, C^Y_m) + 1/NY_m \\ = \tfrac{1}{N} \textstyle\sum_{j \neq m} E^* E^*(Y_j \mid W, A) \mid Y_m, C^Y_m) + 1/NY_m \\ = \tfrac{1}{N} \textstyle\sum_{j \neq m} E^*(\bar{Q}(C^Y_j(W,A)) \mid Y_m, C^Y_m) + 1/NY_m,$$

and

$$P(W,A \mid Y_m, C^Y_m) = I(c^Y_m(W,A) = C^Y_m) \tfrac{P(W,A,Y_m)}{P(Y_m, C^Y_m)} \\ = I(c^Y_m(W,A) = C^Y_m) \tfrac{P(Y_m \mid W,A) P(W,A))}{P(Y_m \mid C^Y_m) P(C^Y_m)} \\ = I(c^Y_m(W,A) = C^Y_m) \tfrac{P(Y_m \mid C^Y_m) P(W,A))}{P(Y_m \mid C^Y_m) P(C^Y_m)} \\ = P(W,A \mid C^Y_m),$$

and thereby

$$E^*(Y^c \mid Y_m, C^Y_m) = \frac{1}{N} \sum_{j \neq m} E^*(\bar{Q}(c^Y_j(W,A)) \mid C^Y_m) + 1/NY_m.$$

Thus,

$$E^*(Y^c \mid C^Y_m) = \frac{1}{N} \sum_{j \neq m} E^*(\bar{Q}(c^Y_j(W,A)) \mid C^Y_m) + 1/N\bar{Q}(C^Y_m).$$

Therefore,

$$\left\{E_{Q,g^*}(Y^c \mid Y(m) = Y(j), C^Y_m = C^Y_j) - E_{Q,g^*}(Y^c \mid C^Y_m = C^Y_j)\right\} = \tfrac{1}{N}\{Y_j - \bar{Q}(C^Y_j)\},$$

which does thus not depend on m.

This proves,

$$\begin{array}{l} D^*(Q,g) = \sum_{j=1}^N \{E_{Q,g^*}(Y^c \mid L(0,j)) - E_{Q,g^*}Y^c\} \\ \qquad + \frac{1}{N}\sum_{j=1}^N \frac{\bar{h}^*}{\bar{h}}(C_j^Y)\left\{Y_j - \bar{Q}(C_j^Y)\right\}. \end{array}$$

Theorem 4 *The efficient influence curve $D^*(P)$ at $P \in \mathcal{M}$ of target parameter $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ is given by*

$$D^*(P) = \sum_{i=1}^N \left\{D^*_{W_i}(Q_W, \bar{Q})(W_i) + \frac{1}{N}\frac{\bar{h}(g^*,Q_W)(C_i^Y)}{\bar{h}(g,Q_W)(C_i^Y)}(Y_i - \bar{Q}(C_i^Y))\right\}$$

where

$$\begin{array}{l} D^*_{W_i}(W_i) = E_{Q,g^*}(Y^c \mid W_i) - E_{Q,g^*}(Y^c) \\ = \frac{1}{N}\sum_{j=1}^N \int_{a,w_{-i}} g^*(a \mid w_{-i}, W_i)\bar{Q}(c_j^Y(a, w_{-i}, W_i) \prod_{l \neq i} Q_{W_l}(w_l) - \psi \\ = \frac{1}{N}\sum_{j=1}^N \{E(Y_j^{g^*} \mid W_i) - E_{W_i}E(Y_j^{g^*} \mid W_i)\}, \end{array}$$

and

$$h_i(g, Q_W)(c) \equiv \int_{a,w,c_i^Y(a,w)=c} g(a \mid w) \prod_{l=1}^N Q_W(w_l) = E_W g_i(c \mid W),$$

where $g_i(c \mid W = w) = P_0(c_i^Y(A, W) = c \mid W = w)$ is the conditional probability that $c_i^Y(A, W)$ equals c, given $W = w$, which is a probability determined by $g(A \mid W)$. In addition, $\bar{h} = \sum_i h_i$, and $\bar{h}^* = \sum_i h_i^*$ with $h_i^* = h_i(g^*, Q_W)$.

Double robustness of efficient influence curve: *Represent the efficient influence curve as $D^*(\bar{Q}, Q_W, g)$. We have*

$$P_0 D^*(\bar{Q}, Q_{W,0}, g_0) = \psi_0 - \Psi(\bar{Q}, Q_{W,0}).$$

Since the efficient influence curve depends on g_0 only through $\bar{h}(g_0, Q_{W,0})$, we have that if $\bar{h}(g, Q_{W,0}) = \bar{h}(g_0, Q_{W,0})$, then

$$P_0 D^*(\bar{Q}, Q_{W,0}, g) = \psi_0 - \Psi(\bar{Q}, Q_{W,0}).$$

We also have that for all g,

$$P_0 D^*(\bar{Q}_{Y,0}, Q_{W,0}, g) = 0.$$

Explicit proof of double robustness: Even though our general theorem can be applied to this single time-point case and this result follows by noting that the second order term $R(Q, Q_0) = 0$, here we provide an explicit proof of the double robustness for this single time point case. We wish to show that $P_0 D^*(Q_{W,0}, \bar{Q}, g_0) = \psi_0 - \psi$. We have

$$E_0 D^*_{W_i}(Q_{W,0}, \bar{Q})(W_i) = \tfrac{1}{N} \textstyle\sum_{j=1}^{N} \int_{a,w} g^*(a \mid w)\bar{Q}(c_j^Y(a,w))Q_{W,0}(w) - \Psi(\bar{Q}, Q_{W,0})$$
$$= 0.$$

We also have

$$\begin{aligned} &E_0 \textstyle\sum_i D^*_{Y_i}(\bar{Q}, Q_{W,0}, g_0) = \frac{1}{N}\sum_i E_0 \frac{\bar{h}_0^*(C_i^Y)}{\bar{h}_0(C_i^Y)}(Y_i - \bar{Q}(C_i^Y)) \\ &= \tfrac{1}{N}\textstyle\sum_i \int \frac{\bar{h}_0^*(c)}{\bar{h}_0(c)}(\bar{Q}_{Y,0} - \bar{Q})(c) h_{i,0}(c) d\mu(c) \\ &= \tfrac{1}{N}\textstyle\int \bar{h}_0^*(c)(\bar{Q}_{Y,0} - \bar{Q})(c) d\mu(c) \\ &= \psi_0 - \Psi(\bar{Q}, Q_{W,0}). \end{aligned}$$

This proves that with $D_i^* = D^*_{W_i} + D^*_{Y_i}$, $D^* = \sum_i D_i^*$, we have

$$E_0 \sum_i D_i^*(Q_{W,0}, \bar{Q}, g_0) = 0 + \psi_0 - \Psi(\bar{Q}, Q_{W,0}).$$

This proves the robustness w.r.t. misspecification of $\bar{Q}$. In addition, $E_0 D^*(\bar{Q}_{Y,0}, Q_{W,0}, g_0) = 0$. □

9.3 Estimating equation approach fails.

Consider the efficient influence curve and let's represent it as an estimating function in ψ:

$$D^*(Q, g, \psi) = \sum_{i=1}^{N} (D^*_{W_i}(Q) - \psi) + D^*_{Y_i}(Q, g_0),$$

where now $D^*_{W_i} = E_Q(Y^{c,g^*} \mid W_i)$. We will represent it as $D^*(Q, h, \psi)$. Given an estimator h_N and Q_N of h_0 and Q_0, respectively, based on the data O, we can estimate ψ with the solution of

$$0 = D^*(Q_N, h_N, \psi)(O).$$

Since $D^*(Q, h, \psi) = D^*(Q, h) - \psi$, this solution is given by

$$\psi_N = D^*(Q_N, h_N)(O) = \sum_{i=1}^{N} D^*_{W_i}(Q_N) + D^*_{Y_i}(Q_N, h_N).$$

Interestingly, this estimator would not be consistent if Q_N is inconsistent and h_N is consistent. This follows since we cannot apply our robustness result if ψ_N is not a substitution estimator $\Psi(Q_N)$, and there is no reason why ψ_N would be a substitution estimator. Indeed, we have $E_0 D^*(Q, h_0, \psi) = N(\Psi(Q) - \psi) + (\psi_0 - \Psi(Q))$ demonstrating that solving $E_0 D^*(Q, h_0, \psi) = 0$ does not imply $\psi = \psi_0$. On the other hand, if $\psi = \Psi(Q)$, then the contribution of $D^*_{W_i}$ equals zero, and we end up with only the contribution from the $D^*_{Y_i}$-components, giving the desired $\psi_0 - \Psi(Q)$. This demonstrates that defining the estimator as a solution of an estimating equation implied by this efficient influence curve fails, while TMLE is able to exploit the *actual* robustness of the efficient influence curve as a function of Q, g, instead of a robustness of a particular estimating function representation of the efficient influence curve.

Remark regarding balancing the two contributions in the efficient influence curve The factor $1/N$ in $D^*_{Y_i}$ might come as a surprise in relation to $D^*_{W_i}$, which is also the factor that completely throws of the estimating equation approach as illustrated above. To intuitively understand that this efficient influence curve does indeed represent a balance between these two contributions, we note the following:

$$\begin{aligned} &\textstyle\sum_i D^*_{W_i} = \sum_i \{E_Q(Y^{c,*} \mid W_i) - \Psi(Q)\} \\ &= \textstyle\sum_{i=1}^N \left\{\frac{1}{N}\sum_{j=1}^N E_Q(Y^*_j \mid W_i) - \Psi(Q)\right\} \\ &= \textstyle\frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N \left\{I(j \in F_i)(E_Q(Y^*_j \mid W_i) - \Psi_j(Q))\right\}, \end{aligned}$$

where $\Psi(Q) = 1/N \sum_j \Psi_j(Q)$, and $\Psi_j(Q) = E_Q Y^*_j$. Thus, indeed the contribution $\sum_i D^*_{W_i}$ is of the same size as function of N as $\sum_i D^*_{Y_i}$, under the assumption that $\mid F_i \mid \leq K < \infty$ for some $K < \infty$, which is indeed an assumption we made to establish $\sqrt{N}$-asymptotics.

9.4 TMLE

Let g^* be given. The target parameter is given by

$$\begin{aligned} \psi_0 &= E_0 Y^{c,g^*} = \Psi(\bar{Q}_{Y,0}, Q_{W,0}) \\ &= \textstyle\frac{1}{N}\sum_{j=1}^N \int_{a,w} \bar{Q}_{Y,0}(c^Y_j(a,w)) g^*(a \mid w) Q_{W,0}(w). \end{aligned}$$

Let $\bar{Q}_N$ be an estimator of $\bar{Q}_{Y,0}(c) = E_0(Y(i) \mid C^Y_i = c)$. Suppose $Y(i) \in \{0,1\}$ or continuous in $(0,1)$. This estimator $\bar{Q}_N$ could be based on the log-likelihood loss function

$$-L(\bar{Q})(O) = \sum_{i=1}^N \log \bar{Q}_Y(c^Y_i)^{Y_i}(1 - \bar{Q}(c^Y_i))^{1-Y_i}.$$

For example, it could be super-learner based on this loss-function. The estimator could also be based on a squared error loss function

$$L_2(\bar{Q})(O) = \sum_{i=1}^{N} (Y_i - \bar{Q}(c_i^Y))^2.$$

Let $\bar{Q}_{W,N}$ be a nonparametric maximum likelihood estimator of Q_W, thus respecting the model for the joint distribution of $W_1, \ldots, W_N$. For example, if W_i are i.i.d., then we would estimate this marginal distribution of W_i with the empirical distribution of $(W_1, \ldots, W_N)$. If $W_1, \ldots, W_N$ are only known to be independent, then we would estimate each marginal distribution of W_i with the discrete distribution that puts mass 1 on the singleton W_i, $i = 1, \ldots, N$: note that this empirical distribution is equivalent with the joint distribution that puts mass 1 on $(W_1, \ldots, W_N)$. A plug-in estimator could now be defined as $\Psi(\bar{Q}_N, Q_{W,N})$.

Let g_N be an estimator of g_0. Given the model assumption $g(A \mid W) = \prod_i g(A_i \mid c_i^A(W))$ for a common conditional density g, this estimator can be based on the log-likelihood loss:

$$L(g)(O) = -\sum_{i=1}^{N} \log g(A_i \mid c_i^A).$$

Given g_N, $Q_{W,N}$, $\bar{Q}_N$, let $\bar{Q}_N(\epsilon)$ be a target-parameter specific submodel through $\bar{Q}_N$ defined by

$$\text{Logit}\bar{Q}_N(\epsilon) = \text{Logit}\bar{Q}_N + \epsilon \frac{\bar{h}(g^*, Q_{W,N})}{\bar{h}(g_N, Q_{W,N})}.$$

Let

$$\epsilon^N = \arg\min_{\epsilon} L(\bar{Q}_N(\epsilon))(O)$$

be the maximum likelihood estimator, which simply involves running univariate logistic regression on a pooled data set with binary outcomes Y_i and covariate $\frac{\bar{h}(g^*, Q_{W,N})}{\bar{h}(g_N, Q_{W,N})}(c_i^Y)$, using as off-set Logit$\bar{Q}_N$. This defines now an update $\bar{Q}_N^* = \bar{Q}_N(\epsilon^N)$. The TMLE of ψ_0 is defined as the corresponding plug-in estimator

$$\psi_N^* = \Psi(\bar{Q}_N^*, Q_{W,N}).$$

We note that this TMLE solves the efficient influence curve equation

$$D^*(\bar{Q}_N^*, Q_{W,N}, g_N, \psi_N^*)(O) = 0.$$

Specifically, by being a substitution estimator $\Psi(Q_N^*)$ and using an NPMLE of $Q_{W,0}$ we have $\sum_i D_{W_i}^*(Q_N^*) = 0$, while the targeted updating of $\bar{Q}_N$ guarantees that $\sum_i D_{Y_i}^*(\bar{Q}_N^*, Q_{W,N}, g_N) = 0$.

10 Asymptotics of TMLE of causal effect of single time point intervention.

In this section we state a theorem establishing the asymptotics of the TMLE of ψ_0 under conditions. Subsequently, we will discuss the theorem shortly and discuss statistical inference in terms of confidence intervals. The proof is deferred to the Appendix. At the end of the Appendix we demonstrate that our proof is generalizable to the general longitudinal data structures.

Theorem 5 *Consider the statistical formulation of data $O = (O_1, \ldots, O_N) \sim P_0 \in \mathcal{M}$, statistical model $\mathcal{M}$, and statistical target parameter $\Psi : \mathcal{M} \to \mathbb{R}$, defined conditionally on the network-profile $\mathbf{F} = (F_1, \ldots, F_N)$. Recall that this network $\mathbf{F}$ implies that Y_i only depends on (W, A) through $(W_j, A_j : j \in F_i)$, and that A_i depends on W through $(W_j : j \in F_i)$. Suppose g_0 is known, and that $W_1, \ldots, W_N$ are only known to be independent. A probability distribution of O is thus parameterized as*

$$P(O) = \prod_{i=1}^{N} P_{W_i}(W_i) g_{0,A_i|W}(A_i \mid W) P_{Y|C^Y}(Y_i \mid C_i^Y), \qquad (6)$$

where $C_i^Y = C_i^Y(A, W) \in \mathcal{C}^Y \subset \mathbb{R}^d$, $P_{Y|C^Y}(\cdot \mid c)$ is a density for Y for each possible $c \in \mathcal{C}^Y$, but is otherwise unspecified, $(g_{0,A_i|W} : i = 1, \ldots, N)$ is known, and each of the marginal distributions P_{W_i} is unspecified. This defines the statistical model $\mathcal{M}$. For a specified stochastic intervention, the target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is defined by

$$\begin{aligned} &\Psi(P) = E_P Y^{c,g^*} = \Psi(\bar{Q}_{Y,0}, Q_{W,0}) \\ &= \tfrac{1}{N} \textstyle\sum_{j=1}^{N} \int_{a,w} \bar{Q}_{Y,0}(c_j^Y(a, w)) g^*(a \mid w) Q_{W,0}(w), \end{aligned}$$

where $Q_{W,0}(w) = P_0(W = w)$, and $\bar{Q}_{Y,0}(c_j^Y(A, W)) = E_0(Y_j \mid A, W)$. Let $D^*(\bar{Q}, Q_W, g_0)(O)$ be the efficient influence curve of Ψ as defined in Theorem 4:

$$D^*(\bar{Q}, Q_W, g_0) = \textstyle\sum_{i=1}^{N} \{D_{W_i}^*(Q_W, \bar{Q})(W_i) + D_{Y_i}^*(\bar{Q}, Q_W)\},$$

where

$$D_{Y,i}^*(\bar{Q}, Q_W) = \frac{1}{N} \frac{\bar{h}(g^*, Q_W)(C_i^Y)}{\bar{h}(g, Q_W)(C_i^Y)} (Y_i - \bar{Q}(C_i^Y)),$$

and $D_{W_i}^* = E(Y^{c,*} \mid W_i) - \Psi(P)$.

Let $Q_{W,N}$ be the distribution that puts mass 1 on $(W_1, \ldots, W_N)$. Consider the TMLE $\psi_N^* = \Psi(Q_N^*) = \Psi(\bar{Q}_N^*, Q_{W,N})$ defined above using the known g_0 in

$\bar{h}(g_0, Q_{W,N})$. *As shown above, this TMLE solves*

$$D^*(\bar{Q}_N^*, Q_{W,N}, g_0)(O) = 0.$$

Recall the definitions of $\bar{h}_0(c) = \sum_{i=1}^N h_{0,i}(c)$, $\bar{h}_0^* = \sum_{i=1}^N h_{0,i}^*$, $h_{0,i}(c) = P_{g_0,Q_{W,0}}(C_i^Y(A,W) = c)$, $h_{0,i}^*(c) = P_{g^*,Q_{W,0}}(C_i^Y(A,W) = c)$, *defined as densities w.r.t. dominating measure* μ, *and let* $\tilde{h}_0 = \bar{h}_0^*/\bar{h}_0$. *Also recall the plug-in estimator* $\tilde{h}_N$ *of* $\tilde{h}_0$ *implied by* g_0 *and* $Q_{W,N}$.

We make the following assumptions:

Entropy condition: *Consider a class* $\mathcal{F}_Y$ *of functions* $c^Y \rightarrow \bar{Q}(c^Y)$ *on a set in* $\mathcal{C}^Y \subset \mathbb{R}^d$ *that contains* $c^Y(A,W)$ *with probability 1. Assume that* $\bar{Q}_N^* \in \mathcal{F}_Y$ *with probability 1. Consider a class* $\mathcal{F}_h$ *of functions* $c^Y \rightarrow \bar{h}(c^Y)$ *on* $\mathcal{C}^Y \subset \mathbb{R}^d$. *Assume that* $\tilde{h}_N^* \in \mathcal{F}_h$ *with probability 1. Define the dissimilarity measure on the cartesian product of* $\mathcal{F} = \mathcal{F}_Y \times \mathcal{F}_h$*:*

$$d((\tilde{h}_1, \bar{Q}_1), (\tilde{h}, \bar{Q})) = \max\left(\sup_{c \in \mathcal{C}^Y} \mid \tilde{h}_1 - \tilde{h} \mid, \sup_{c \in \mathcal{C}^Y} \mid \bar{Q}_1 - \bar{Q} \mid\right).$$

Assume that there exists some $\eta > 0$, *so that* $\int_0^\eta \sqrt{\log(N(\epsilon, \mathcal{F}, d))} d\epsilon < \infty$.

In particular, this assumption holds if $\sup_{f \in \mathcal{F}_Y} \parallel f \parallel_v^* < \infty$ *and* $\sup_{f \in \mathcal{F}_h} \parallel f \parallel_v^* < \infty$, *where* $\parallel f \parallel_v^*$ *is the uniform sectional variation norm as defined in Gill, van der Laan, Wellner (1996) and van der Laan (1996).*

Universal bound: *Assume* $\sup_{f \in \mathcal{F}, O} \mid f \mid (O) < \infty$, *where the supremum of* O *is over a set that contains* O *with probability* 1. *This assumption will typically be a consequence of the entropy condition, such as it is a consequence of the uniform sectional variation norm condition above.*

Consistency and rate condition: *Assume* $d(\tilde{h}_N, \bar{Q}_N^*), (\tilde{h}_0, \bar{Q}^*)) \rightarrow 0$ *in probability as* $N \rightarrow \infty$,

$$R_{N,1} \equiv -P_0 \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(\bar{Q}_N^* - \bar{Q}_Y^*) = o_P(1/\sqrt{N})$$

and

$$R_{N,4} = \int_c \left\{\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}}\right\} \frac{1}{\bar{h}_0} (\bar{h}_N - \bar{h}_0)(\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c)\bar{h}_0 d\mu = o_P\left(\frac{1}{\sqrt{N}}\right).$$

Positivity type condition: *Assume*

$$\sup_{c\in\mathcal{C}^Y}\frac{\bar{h}^*(g^*,Q_{W,0})}{\bar{h}(g_0,Q_{W,0})}(c)<\infty.$$

Network structure condition: *Assume that there exists a* $K<\infty$ *so that* $\sup_i \mid F_i \mid < K$ *for all* $i=1,\ldots,$ *a.s.*

Restriction on dependence of stochastic intervention: *Assume that* $g^*((A_j : j\in F_i)\mid W)$ *only depends on* W *through* $(W_j : j\in S_i)$ *for sets* S_i *deterministically implied by* $\mathbf{F}$ *with* $\sup_i \mid S_i \mid < K < \infty$ *for some* $K<\infty$.

First order approximation: *Then,*

$$\psi_N^*-\psi_0=\frac{1}{N}\sum_{i=1}^N\{f_i(O)-P_0f_i\}+o_P(1/\sqrt{N}),$$

where

$$\begin{aligned}
f_i &= D^*_{Y,i}(\bar{Q}^*,Q_{W,0})+f^1_{W,i}+f^2_{W,i}\\
f^1_{W,i}(W) &= \int_a \bar{Q}^*_Y(c^Y_i(a,W))g^*(a\mid W)\\
f^2_{W,i}(W) &= \int_c\left\{\frac{h^*_{i,N}}{\bar{h}_0}-\frac{\bar{h}^*_0}{\bar{h}^2_0}h_{i,N}\right\}(c)(\bar{Q}_{Y,0}-\bar{Q}^*_Y)(c)\bar{h}_0(c)\\
h^*_{i,N}(c) &= \int_a I(c^Y_i(a,W)=c)g^*(a\mid W)=g^*_i(c\mid W)\\
h_{i,N}(c) &= \int_a I(c^Y_i(a,W)=c)g_0(a\mid W)=g_{0,i}(c\mid W).
\end{aligned}$$

Weak convergence of first order approximation: *We can orthogonally decompose*

$$f_i(O)-P_0f_i=f_{Y,i}(O)+f_{A,i}(O)+f_{W,i}(O),$$

where

$$\begin{aligned}
f_{Y,i} &= D^*_{Y,i} - E_0(D^*_{Y,i} \mid A, W) \\
&= \frac{\bar{h}^*_0}{\bar{h}_0}(C^Y_i)(Y_i - \bar{Q}_{Y,0}(C^Y_i)) \\
f_{A,i} &= E_0(D^*_{Y,i} \mid A, W) - E_0(D^*_{Y,i} \mid W) \\
&= \frac{\bar{h}^*_0}{\bar{h}_0}(C^Y_i)(\bar{Q}_{Y,0} - \bar{Q}^*_Y)(C^Y_i) \\
&\quad - \int_c \frac{\bar{h}^*_0}{\bar{h}_0}(c)(\bar{Q}_{Y,0} - \bar{Q}^*_Y)(c) g_{0,i}(c \mid W) \\
E_0(D^*_{Y,i} \mid W) &= \int_c \frac{\bar{h}^*_0}{\bar{h}_0}(\bar{Q}_{Y,0} - \bar{Q}^*_Y)(c) g_{0,i}(c \mid W) \\
f_{W,i} &= f^1_{W,i} + f^2_{W,i} + E_0(D^*_{Y,i} \mid W) - P_0\{f^1_{W,i} + f^2_{W,i} + E_0(D^*_{Y,i} \mid W)\} \\
&= \int_a \bar{Q}_{Y,0}(c^Y_i(a, W)) g^*(a \mid W) - \int_a \bar{Q}_{Y,0}(c^Y_i(a, W)) g^*(a \mid w) Q_{W,0}(w) \\
&= \int_c \bar{Q}_{Y,0}(c) g^*_i(c \mid W) - \int_{c,w} \bar{Q}_{Y,0}(c) g^*_i(c \mid W) Q_{W,0}(w).
\end{aligned}$$

For $(i, j) \in \{1, \ldots, N\}^2$*, define* $R_W(i, j) = I(S_i \cap S_j \neq \emptyset)$*,* $R_A(i, j) = I(F_i \cap F_j \neq \emptyset)$*, and* $R_2(i, j) = I(R_A(i, j) = 1 \text{ or } R_W(i, j) = 1)$*. We have*

$$\frac{1}{\sqrt{N}} \sum_i \{f_i(O) - P_0 f_i\} \Rightarrow_d N(0, \sigma^2), \text{ where } \sigma^2 = \sigma^2_Y + \sigma^2_A + \sigma^2_W,$$

and

$$\sigma^2_Y = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^N P_0 f^2_{Y,i}$$

$$\sigma^2_A = \lim_{N \to \infty} \frac{1}{N} \sum_{i_1, i_2} R_A(i_1, i_2) P_0 f_{A,i_1} f_{A,i_2}$$

$$\sigma^2_W = \lim_{N \to \infty} \frac{1}{N} \sum_{i_1, i_2} R_W(i_1, i_2) P_0 f_{W,i_1} f_{W,i_2},$$

and $P_0 f$ *denotes the marginal expectation of* $f(O)$*, given* $\mathbf{F}$*. As a consequence,* $\sqrt{N}(\psi^*_N - \psi_0) \Rightarrow_d N(0, \sigma^2)$*.*

Alternative expression of asymptotic variance: *One can also represent* σ^2 *as*

$$\sigma^2 = \lim_{N \to \infty} \frac{1}{N} \sum_{i_1, i_2} R_2(i_1, i_2) P_0 f_{i_1} f_{i_2}.$$

10.1 Statistical inference

One can estimate σ^2 by plugging in estimators $Q_{W,N}, \bar{Q}_N^*$ in the expressions for $f_{Y,i}, f_{W,i}, f_{A,i}$. Given an estimator of σ_N^2, one can then construct a confidence interval $\psi_N^* \pm 1.96\sigma_N/\sqrt{N}$. If σ_N is consistent for σ, then this will be an asymptotically valid 0.95-confidence interval. The expression for σ^2 suggests that a consistent estimator of σ^2 relies on consistent estimation of $\bar{Q}_{Y,0}$, even though the consistency of ψ_N^* only relies on a consistent estimator of h_0 and thus g_0 (since the expectation w.r.t. W is consistently estimated). Therefore, even if ψ_N^* relied on a less nonparametric estimator of $\bar{Q}_{Y,0}$, we recommend using a super learner using flexible machine learning algorithms when estimating this asymptotic variance σ^2.

On the other hand, we know that in the i.i.d. setting, one can obtain a consistent estimator of σ^2 only relying on a consistent estimator of g_0, thereby allowing that $\bar{Q}_N^*$ is inconsistent. It remains to be investigated till what degree such robust (possibly conservative) estimation of σ^2 is possible in greater generality.

We claim that if g_0 is unknown, and one uses an MLE g_N according to some model, then σ^2 is a lower bound for the actual asymptotic variance of the TMLE, based on a generalization of the result in van der Laan and Robins (2003). This is due to the fact that g_0 is an orthogonal nuisance parameter. Such a result would then allow us to use this same plug-in estimator σ_N^2 (using g_N for g_0) in the statistical model $\mathcal{M}$ in which g_0 is not known but a correctly specified model for g_0 is available. Again, such a result will need to be formally established in future research.

11 TMLE of Causal Effect of Single Time Point Intervention, when weakening the sequential conditional independence assumption

One of the fundamental assumptions of our general causal model was that the U_i, $i = 1, \ldots, N$, were independent, which implied that $L_i(t), A_i(t)$, $i = 1, \ldots, N$, are conditionally independent, given $\bar{L}(t-1), \bar{A}(t-1)$, for all $t = 0, \ldots, \tau + 1$. The latter property we referred to as the sequential conditional independence assumption, and it implied our factorized likelihood, and allowed us to establish a particular central limit theorem for our TMLE. We might want to weaken this assumption by only assuming that $(L_i(t), A_i(t))$ and $(L_j(t), A_j(t))$ are conditionally independent, given $\bar{L}(t-1), \bar{A}(t-1)$, if

$F_i(t) \cap F_j(t) = \emptyset$. Returning to our $\tau = 0$-case, this corresponds with assuming that U_i and U_j are independent if $F_i \cap F_j = \emptyset$. Specifically, this corresponds with assuming that, if $F_i \cap F_j = \emptyset$, then W_i and W_j are independent, A_i and A_j are independent, given W, and Y_i and Y_j are independent, given A, W. As a consequence, under this weaker assumption, the likelihood is given by:

$$P(O) = Q_W(W) g_A(A \mid W) Q_Y(Y \mid A, W),$$

where the joint distribution Q_W, the joint conditional distribution g_A and Q_Y satisfy these conditional independence assumption implied by the network $(F_i : i = 1, \ldots, N)$. We still assume the randomization assumption stating that A is independent of Y_{g^*}, given W. However, since Y_i is affected by (W_j, A_j), $j \in F_i$, and A_j is confounded by variables $(W_l : l \in F_j)$ that have arrows towards Y_i through Y_j, it does not suffice anymore to only adjust for $(W_j, A_j : j \in F_i)$ when identifying EY_{i,g^*}. That is, to identify the counterfactual mean EY_{i,g^*} we need to adjust for $(W_l : l \in R_i)$, where $R_i = \cup_{j \in F_i} F_j$. Let $c_i^Y(a, w)$ be a summary measure of $(A_j : j \in F_i)$ and $(W_j : j \in R_i)$, and assume that there exists a common function $\bar{Q}$ so that $E(Y_i \mid A, W) = \bar{Q}(c_i^Y(A, W))$, $i = 1, \ldots, N$. Thus, we assume a model on the conditional distributions of Y_i, given A, W, for each $i = 1, \ldots, N$, but avoid modeling of the actual joint distribution of the N-dimensional outcome Y beyond the stated conditional independence assumption. This defines now the statistical model $\mathcal{M}$ for the distribution of O, which is more nonparametric than our previous model covered in the previous section relying on the sequential conditional independence assumption.

We have the following identifiability result for the counterfactual mean $E_0 Y_{g^*}^c = E_0(1/N \sum_i Y_{i,g^*})$:

$$EY_{g^*}^c = \frac{1}{N} \sum_{i=1}^{N} \int_{w,a} Q_W(w) g^*(a \mid w) \bar{Q}(c_i^Y(a, w)) \equiv \Psi(P).$$

This defines now also our statistical target parameter $\Psi : \mathcal{M} \rightarrow \mathbb{R}$. Consider the following function:

$$D^*(P) = D_W^*(Q_W, \bar{Q})(W) + \sum_{i=1}^{N} D_{Y_i}^*(\bar{Q}, Q_W, g)(O),$$

where

$$\begin{array}{l} D_{Y_i}^*(Q, g)(O) = 1/N \frac{\bar{h}(g^*, Q_W)}{\bar{h}(g, Q_W)} (Y_i - \bar{Q}(C_i^Y) \\ D_W^*(Q_W, \bar{Q})(W) = 1/N \sum_i \int_a g^*(a \mid W) \bar{Q}(c_i^Y(a, W)) - \Psi(Q) \\ \bar{h}(g, Q_W)(c) = 1/N \sum_{i=1}^N P_{Q,g}(c_i^Y(A, W) = c) \\ \bar{h}(g^*, Q_W)(c) = 1/N \sum_{i=1}^N P_{Q,g^*}(c_i^Y(A, W) = c). \end{array}$$

We have the same double robustness results as before: $E_0 D^*(Q_{W,0},\bar{Q},g_0) = \psi_0 - \Psi(Q_{W,0},\bar{Q})$ and $E_0 D^*(Q_0,g) = 0$ for all g. This is shown as before: Firstly,

$$E_0 D^*_W(Q_{W,0},\bar{Q})(W_i) = \tfrac{1}{N}\textstyle\sum_{j=1}^N \int_{a,w} g^*(a \mid w)\bar{Q}(c_j^Y(a,w))Q_{W,0}(w) - \Psi(\bar{Q},Q_{W,0}) \\ = 0.$$

We also have

$$\begin{array}{l} E_0 \sum_i D^*_{Y_i}(\bar{Q},Q_{W,0},g_0) = \frac{1}{N}\sum_i E_0 \frac{\bar{h}_0^*(C_i^Y)}{\bar{h}_0(C_i^Y)}(Y_i - \bar{Q}(C_i^Y)) \\ = \frac{1}{N}\sum_i \int \frac{\bar{h}_0^*(c)}{\bar{h}_0(c)}(\bar{Q}_{Y,0} - \bar{Q})(c)h_{i,0}(c)d\mu(c) \\ = \frac{1}{N}\int \bar{h}_0^*(c)(\bar{Q}_{Y,0} - \bar{Q})(c)d\mu(c) \\ = \psi_0 - \Psi(\bar{Q},Q_{W,0}). \end{array}$$

This proves the robustness w.r.t. misspecification of $\bar{Q}$:

$$E_0 \sum_i D_i^*(Q_{W,0},\bar{Q},g_0) = 0 + \psi_0 - \Psi(\bar{Q},Q_{W,0}).$$

In addition, it follows trivially that $E_0 D^*(\bar{Q}_{Y,0},Q_{W,0},g_0) = 0$. Due to this fact that our previous efficient influence curve is still a valid estimating function satisfying the same double robustness, we propose to use the same TMLE procedure to estimate ψ_0. It might still be of interest to determine the actual efficient influence curve and a corresponding TMLE, but we leave that for possible future research.

11.1 TMLE

Let $\bar{Q}_N$ be an estimator of $\bar{Q}_{Y,0}(c) = E_0(Y_i \mid C_i^Y = c)$. Suppose Y_i is binary in $\{0,1\}$ or continuous with values in $(0,1)$. This estimator $\bar{Q}_N$ could be based on the log-likelihood loss function

$$-L(\bar{Q})(O) = \sum_{i=1}^N \log \bar{Q}_Y(c_i^Y)^{Y_i}(1-\bar{Q}(c_i^Y))^{1-Y_i}.$$

For example, it could be a super learner based on this loss function. The estimator could also be based on a squared error loss function

$$L_2(\bar{Q})(O) = \sum_{i=1}^N (Y_i - \bar{Q}(c_i^Y))^2.$$

Let $\bar{Q}_{W,N}$ be the empirical distribution that puts mass 1 on $W = (W_1, \ldots, W_N)$. A plug-in estimator could now be defined as $\Psi(\bar{Q}_N, Q_{W,N})$.

Let g_N be an estimator of g_0, respecting the statistical model for g_0. For example, one could factorize $g_0(A \mid W) = \prod_{i=1}^N g_{0,i}(A_i \mid A_1, \ldots, A_{i-1}, W)$, and utilizing the known conditional independencies of A_i with A_j, given W, and assume a common model $g_{0,i}(A_i \mid A_1, \ldots, A_{i-1}, W) = g_0(A_i \mid C_i^A)$ for some summary measure C_i^A.

Given g_N, $Q_{W,N}$, $\bar{Q}_N$, let $\bar{Q}_N(\epsilon)$ be a target-parameter specific submodel through $\bar{Q}_N$ defined by

$$\text{Logit}\bar{Q}_N(\epsilon) = \text{Logit}\bar{Q}_N + \epsilon \frac{\bar{h}(g^*, Q_{W,N})}{\bar{h}(g_N, Q_{W,N})}.$$

Let

$$\epsilon^N = \arg\min_{\epsilon} L(\bar{Q}_N(\epsilon))(O)$$

be the maximum likelihood estimator, which simply involves running univariate logistic regression on a pooled data set with binary outcomes Y_i and covariate $\frac{\bar{h}(g^*,Q_{W,N})}{\bar{h}(g_N,Q_{W,N})}(c_i^Y)$, using as off-set Logit$\bar{Q}_N$. This defines now an update $\bar{Q}_N^* = \bar{Q}_N(\epsilon^N)$. The TMLE of ψ_0 is defined as the corresponding plug-in estimator

$$\psi_N^* = \Psi(\bar{Q}_N^*, Q_{W,N}).$$

We note that this TMLE solves the equation

$$D^*(\bar{Q}_N^*, Q_{W,N}, g_N, \psi_N^*)(O) = 0.$$

Specifically, by being a substitution estimator $\Psi(Q_N^*)$ and using an NPMLE of $Q_{W,0}$ we have $D_W^*(Q_{W,N}, \bar{Q}_N^*)(O) = 0$, while the targeted updating of $\bar{Q}_N$ guarantees that $\sum_i D_{Y_i}^*(\bar{Q}_N^*, Q_{W,N}, g_N) = 0$ as well.

Since the outcomes are not conditionally independent, given W, A, the statistical properties of cross-validation to select among candidate estimators of $\bar{Q}_0$ are not well understood yet. For example, would it be appropriate to split the N observations randomly into a training and validation sample, ignoring these dependencies (since it is not obvious how to divide up the sample in an independent training and validation sample without discarding lots of observations)? The same remark applies to the use of cross-validation to select among candidate estimators of the conditional distribution of A, given W. Here we suffice with the following remark. In order to study the properties of such a cross-validation selector we will need to understand empirical processes of the form

$$Z_N(f) = \sum_{i=1}^N f_i(f)(O),$$

for a specified mapping $f \rightarrow f_i(f)$ from a class of functions $\mathcal{F}$ into functions of O. With each i there is an associated set $R_i \subset \{1,\ldots,N\}$, and it is known that $f_i(f)$ is independent of $f_j(f)$ if $R_i \cap R_j = \emptyset$. In the appendix we provide general conditions under which this process $Z_N = (Z_N(f) : f \in \mathcal{F})$ converges weakly to a Gaussian process $Z = (Z(f) : f \in \mathcal{F})$. For completeness, we provide here the general theorem that is a corollary from the Appendix:

Theorem 6 *Consider a process $Z_N = (Z_N(f) : f \in \mathcal{F})$, with $Z_N(f) = 1/\sqrt{N}\sum_{i=1}^N f_i(f)(O)$, where $E_0 f_i(f)(O) = 0$, $f_i(f)$ is independent of $f_j(f)$ if $R_i \cap R_j = \emptyset$, where $R_i \subset \{1,\ldots,N\}$ is a subset of indices, $i = 1,\ldots,N$, and $\mathcal{F}$ is a set of multivariate uniformly bounded real valued functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$. We make the following assumptions:*

- $\sup_i \mid R_i \mid < K$ *for some universal* $K < \infty$.
- *For all integers $p > 0$, $\{E_0 f_i(f)(O)^p\}^{1/p} \leq C \parallel f \parallel_\infty$ for supremum norm $\parallel f \parallel$ on $\mathcal{F}$, and universal $C < \infty$.*
- *The entropy integral $\int \sqrt{\log N(\epsilon, \mathcal{F}, \parallel \cdot \parallel_\infty)} d\epsilon < \infty$ for $\mathcal{F}$ w.r.t. norm $\parallel \cdot \parallel_\infty$ is finite.*
- *The marginal distributions $Z_N(f)$ converge to a normal distribution $Z(f)$ for all $f \in \mathcal{F}$.*

Then Z_N converges weakly to a Gaussian process Z identified by the covariance operator $\Sigma(f_1, f_2)$ defined by

$$\Sigma(f_1, f_2) = \lim_{N\rightarrow\infty} \tfrac{1}{N} \textstyle\sum_{i=1}^N \sum_{j=1}^N I(R_i \cap R_j \neq \emptyset) E_0 f_i(f_1) f_j(f_2).$$

We conjecture that, using these fundamental empirical process results, it will be possible to establish certain types of oracle inequalities for the cross-validation selector for $\bar{Q}_0$ that simply ignores the conditional dependencies of the Y_i's, given A, W, between the N observations. Similarly, such results would then also be obtained for more general conditional distributions such as g_0. This will be an area of interest for future research.

11.2 Asymptotic theorem

We can obtain a similar asymptotics theorem for this TMLE as the one presented in the previous section, due to the fact that it is defined as a similar target parameter and uses the same TMLE algorithm. The only difference is that the covariance functions are different due to having allowed for many more dependencies.

Theorem 7 *Consider the statistical formulation of data $O = (O_1, \ldots, O_N) \sim P_0 \in \mathcal{M}$, statistical model $\mathcal{M}$, and statistical target parameter $\Psi : \mathcal{M} \to \mathrm{I\!R}$, defined conditionally on the network-profile $\mathbf{F} = (F_1, \ldots, F_N)$. Suppose g_0 is known. Specifically, the model corresponds with assuming that, if $F_i \cap F_j = \emptyset$, then W_i and W_j are independent, A_i and A_j are independent, given W, and Y_i and Y_j are independent, given A, W. The likelihood is given by:*

$$P(O) = Q_W(W) g_A(A \mid W) Q_Y(Y \mid A, W),$$

where the joint distribution Q_W, the joint conditional distribution g_A and Q_Y satisfy these conditional independence assumption implied by the network $(F_i : i = 1, \ldots, N)$. Let $R_i = \cup_{j \in F_i} F_j$. Let $c_i^Y(a, w)$ be a summary measure of $(A_j : j \in F_i)$ and $(W_j : j \in R_i)$, and assume that there exists a common function $\bar{Q}$ so that $E(Y_i \mid A, W) = \bar{Q}(c_i^Y(A, W))$, $i = 1, \ldots, N$. This defines the statistical model $\mathcal{M}$ for the distribution of O. The statistical target parameter $\Psi : \mathcal{M} \to \mathrm{I\!R}$ is defined as:

$$\Psi(P) = \frac{1}{N} \sum_{i=1}^{N} \int_{w,a} Q_W(w) g^*(a \mid w) \bar{Q}(c_i^Y(a, w)).$$

Consider the following function:

$$D^*(P) = D^*_W(Q_W, \bar{Q})(W) + \sum_{i=1}^{N} D^*_{Y_i}(\bar{Q}, Q_W, g)(O),$$

where

$$\begin{aligned}
&D^*_{Y_i}(Q, g)(O) = 1/N \tfrac{\bar{h}(g^*, Q_W)}{\bar{h}(g, Q_W)} (Y_i - \bar{Q}(C_i^Y) \\
&D^*_W(Q_W, \bar{Q})(W) = 1/N \textstyle\sum_i \int_a g^*(a \mid W) \bar{Q}(c_i^Y(a, W)) - \Psi(Q) \\
&\bar{h}(g, Q_W)(c) = 1/N \textstyle\sum_{i=1}^N P_{Q,g}(c_i^Y(A, W) = c) \\
&\bar{h}(g^*, Q_W)(c) = 1/N \textstyle\sum_{i=1}^N P_{Q,g^*}(c_i^Y(A, W) = c).
\end{aligned}$$

Let $Q_{W,N}$ be the distribution that puts mass 1 on $(W_1, \ldots, W_N)$. Consider the TMLE $\psi^*_N = \Psi(Q^*_N) = \Psi(\bar{Q}^*_N, Q_{W,N})$ defined above using the known g_0 in $\bar{h}(g_0, Q_{W,N})$. As shown above, this TMLE solves

$$D^*(\bar{Q}^*_N, Q_{W,N}, g_0)(O) = 0.$$

Recall the definitions of $\bar{h}_0(c) = \sum_{i=1}^N h_{0,i}(c)$, $\bar{h}^*_0 = \sum_{i=1}^N h^*_{0,i}$, $h_{0,i}(c) = P_{g_0, Q_{W,0}}(C_i^Y(A, W) = c)$, $h^*_{0,i}(c) = P_{g^*, Q_{W,0}}(C_i^Y(A, W) = c)$, defined as densities w.r.t. dominating measure μ, and let $\tilde{h}_0 = \bar{h}^*_0 / \bar{h}_0$. Also recall the plug-in estimator $\tilde{h}_N$ of $\tilde{h}_0$ implied by g_0 and $Q_{W,N}$.

We make the following assumptions:

Entropy condition: *Consider a class $\mathcal{F}_Y$ of functions $c^Y \rightarrow \bar{Q}(c^Y)$ on a set in $\mathcal{C}^Y \subset \mathbb{R}^d$ that contains $c^Y(A,W)$ with probability 1. Assume that $\bar{Q}^*_N \in \mathcal{F}_Y$ with probability 1. Consider a class $\mathcal{F}_h$ of functions $c^Y \rightarrow \bar{h}(c^Y)$ on $\mathcal{C}^Y \subset \mathbb{R}^d$ that contains $c^Y(A,W)$ with probability 1. Assume that $\tilde{h}^*_N \in \mathcal{F}_h$ with probability 1. Define the dissimilarity measure on the cartesian product of $\mathcal{F} = \mathcal{F}_Y \times \mathcal{F}_h$:*

$$d((\tilde{h}_1, \bar{Q}_1), (\tilde{h}, \bar{Q})) = \max \left(\sup_{c \in \mathcal{C}^Y} \mid \tilde{h}_1 - \tilde{h} \mid, \sup_{c \in \mathcal{C}^Y} \mid \bar{Q}_1 - \bar{Q} \mid \right).$$

Assume that there exists some $\eta > 0$, so that $\int_0^\eta \sqrt{\log\left(N(\epsilon, \mathcal{F}, d)\right)} d\epsilon < \infty$.

In particular, this assumption holds if $\sup_{f \in \mathcal{F}_Y} \parallel f \parallel_v^* < \infty$ and $\sup_{f \in \mathcal{F}_h} \parallel f \parallel_v^* < \infty$, where $\parallel f \parallel_v^*$ is the uniform sectional variation norm as defined in Gill, van der Laan, Wellner (1996) and van der Laan (1996).

Universal bound: *Assume $\sup_{f \in \mathcal{F}, O} \mid f \mid (O) < \infty$, where the supremum of O is over a set that contains O with probability* 1. *This assumption will typically be a consequence of the entropy condition, such as it is a consequence of the uniform sectional variation norm condition above.*

Consistency and rate condition: *Assume $d(\tilde{h}_N, \bar{Q}^*_N), (\tilde{h}_0, \bar{Q}^*)) \rightarrow 0$ in probability as $N \rightarrow \infty$,*

$$R_{N,1} \equiv -P_0 \left(\frac{\bar{h}^*_N}{\bar{h}_N} - \frac{\bar{h}^*_0}{\bar{h}_0} \right) (\bar{Q}^*_N - \bar{Q}^*_Y) = o_P(1/\sqrt{N})$$

and

$$R_{N,4} = \int_c \left\{ \frac{\bar{h}^*_N}{\bar{h}_N} - \frac{\bar{h}^*_0}{\bar{h}} \right\} \frac{1}{\bar{h}_0} (\bar{h}_N - \bar{h}_0)(\bar{Q}_{Y,0} - \bar{Q}^*_Y)(c) \bar{h}_0 d\mu(c) = o_P\left(\frac{1}{\sqrt{N}} \right).$$

Positivity type condition: *Assume*

$$\sup_{c \in \mathcal{C}^Y} \frac{\bar{h}^*(g^*, Q_{W,0})}{\bar{h}(g_0, Q_{W,0})}(c) < \infty.$$

Network structure condition: *Assume that there exists a $K < \infty$ so that $\sup_i \mid F_i \mid < K$ for all $i = 1, \ldots,$ a.s.*

Restriction on dependence of stochastic intervention: *Assume that $g^*((A_j : j \in F_i) \mid W)$ only depends on W through $(W_j : j \in S_i)$ for sets $S_i \subset \{1, \ldots, N\}$ with $\mid S_i \mid < K$ for some $K < \infty$.*

First order approximation: *Then,*

$$\psi_N^* - \psi_0 = \frac{1}{N}\sum_{i=1}^{N}\{f_i(O) - P_0 f_i\} + o_P(1/\sqrt{N}),$$

where

$$\begin{aligned}
f_i &= D_{Y,i}^*(\bar{Q}^*, Q_{W,0}) + f_{W,i}^1 + f_{W,i}^2 \\
f_{W,i}^1(W) &= \int_a \bar{Q}_Y^*(c_i^Y(a,W))g^*(a \mid W) \\
f_{W,i}^2(W) &= \int_c \left\{ \frac{h_{i,N}^*}{\bar{h}_0} - \frac{\bar{h}_0^*}{\bar{h}_0^2}h_{i,N} \right\}(c)(\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c)\bar{h}_0(c)d\mu(c) \\
h_{i,N}^*(c) &= \int_a I(c_i^Y(a,W) = c)g^*(a \mid W) = g_i^*(c \mid W) \\
h_{i,N}(c) &= \int_a I(c_i^Y(a,W) = c)g_0(a \mid W) = g_{0,i}(c \mid W).
\end{aligned}$$

Weak convergence of first order approximation: *We can orthogonally decompose*

$$f_i(O) = f_{Y,i}(O) + f_{A,i}(O) + f_{W,i}(O),$$

where

$$\begin{aligned}
f_{Y,i} &= D_{Y,i}^* - E_0(D_{Y,i}^* \mid A, W) \\
&= \frac{\bar{h}_0^*}{\bar{h}_0}(C_i^Y)(Y_i - \bar{Q}_{Y,0}(C_i^Y)) \\
f_{A,i} &= E_0(D_{Y,i}^* \mid A, W) - E_0(D_{Y,i}^* \mid W) \\
&= \frac{\bar{h}_0^*}{\bar{h}_0}(C_i^Y)(\bar{Q}_{Y,0} - \bar{Q}_Y^*)(C_i^Y) \\
&\quad - \int_c \frac{\bar{h}_0^*}{\bar{h}_0}(c)(\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c)g_{0,i}(c \mid W) \\
E_0(D_{Y,i}^* \mid W) &= \int_c \frac{\bar{h}_0^*}{\bar{h}_0}(\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c)g_{0,i}(c \mid W) \\
f_{W,i} &= f_{W,i}^1 + f_{W,i}^2 + E_0(D_{Y,i}^* \mid W) - P_0\{f_{W,i}^1 + f_{W,i}^2 + E_0(D_{Y,i}^* \mid W)\} \\
&= \int_a \bar{Q}_{Y,0}(c_i^Y(a,W))g^*(a \mid W) - \int_a \bar{Q}_{Y,0}(c_i^Y(a,W))g^*(a \mid w)Q_{W,0}(w) \\
&= \int_c \bar{Q}_{Y,0}(c)g_i^*(c \mid W) - \int_{c,w} \bar{Q}_{Y,0}(c)g_i^*(c \mid w)Q_{W,0}(w).
\end{aligned}$$

We make the following observations: 1) $f_{Y,i}$ depends only on Y_i, given A,W, so that $f_{Y,i}, f_{Y,j}$ are independent, given A,W, if $F_i \cap F_j = \emptyset$, 2) $f_{A,i}$ is a function of A_j, $j \in F_i$, so that $f_{A,i}$ and $f_{A,j}$ are independent, given W, if $F_i^+ \cap F_j^+ = \emptyset$, where $F_i^+ = \cup_{l \in F_i} F_l$, 3) $f_{W,i}$ is a function of $(W_j : j \in S_i)$, so that $f_{W,i}$ and $f_{W,j}$ are independent if $S_i^+ \cap S_j^+ = \emptyset$, where $S_i^+ = \cup_{l \in S_i} F_l$. Define $R_Y(i,j) = I(F_i \cap F_j \neq \emptyset)$, $R_A(i,j) = I(F_i^+ \cap F_j^+ \neq \emptyset)$, $R_W(i,j) = I(S_i^+ \cap S_j^+ \neq \emptyset)$ as corresponding indicators of not being independent. Let $R(i,j) = I(R_Y(i,j) = 1$ or $R_A(i,j) = 1$ or $R_W(i,j) = 1)$.

We have

$$\frac{1}{\sqrt{N}} \sum_i \{f_i(O) - P_0 f_i\} \Rightarrow_d N(0,\sigma^2), \text{ where } \sigma^2 = \sigma_Y^2 + \sigma_A^2 + \sigma_W^2,$$

and

$$\begin{aligned} \sigma_Y^2 &= \lim_{N\to\infty} \frac{1}{N} \sum_{i_1,i_2}^{N} R_Y(i_1,i_2) P_0 f_{Y,i_1} f_{Y,i_2} \\ \sigma_A^2 &= \lim_{N\to\infty} \frac{1}{N} \sum_{i_1,i_2} R_A(i_1,i_2) P_0 f_{A,i_1} f_{A,i_2} \\ \sigma_W^2 &= \lim_{N\to\infty} \frac{1}{N} \sum_{i_1,i_2} R_W(i_1,i_2) P_0 f_{W,i_1} f_{W,i_2}, \end{aligned}$$

and $P_0 f$ denotes the marginal expectation of $f(O)$, given $\mathbf{F}$. As a consequence, $\sqrt{N}(\psi_N^* - \psi_0) \Rightarrow_d N(0,\sigma^2)$.

Alternative expression of asymptotic variance: *One can also represent σ^2 as*

$$\sigma^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{i_1,i_2} R(i_1,i_2) P_0 f_{i_1} f_{i_2}.$$

12 When only, but carefully, observing a random sample of the units in the network

We assume the same structural causal model and define the same causal quantity of interest $\psi_0^F = E_0 Y_{g^*}^c$, where $Y_{g^*}^c = 1/N \sum_{i=1}^N Y_{i,g^*}$. We already demonstrated that ψ_0^F can be represented as a function Ψ evaluated at Q_0, where Q_0 represents the collection of conditional distributions of $L_i(t)$, given $C_i^{L(t)}$, across t. However, we now consider the case that we do not observe the data on all N individuals.

Let

$$O_i \;\; = \;\; (L_i(t), c_t^L(Pa(L_i(t))), A_i(t), c_t^A(Pa(A_i(t)))) : t)$$

be the data structure for subject i augmented with the summary measures of the parents of $L_i(t)$ and $A_i(t)$ for all t. Note that the data $O_i(t)$ on subject i at time t potentially partially overlaps with data on another subject j due to the parent-data for subject i at time t that includes information about subject j. We assume that we observe a random sample of size n from $\{1, \ldots, N\}$, and that for each of these individuals we observe this data structure O_i: thus, our observed data is O_i, $i = 1, \ldots, n$.

To summarize, we observe a random sample of size n of the N individuals making up the population and network of interest for which we assumed our posed structural causal model, and for each of the individuals in this random sample we observe all relevant variables that are used by the structural equations for subject i to generate its realized process (L_i, A_i). Just for clarity, this does not mean that if subject i is observed, and subject i is connected to subject j so that O_i includes data on subject j, that we observe all the data on subject j.

Partial likelihood for random sample of individuals: Under our causal model, a partial likelihood of $O = (O_1, \ldots, O_n)$ (i.e., a likelihood that only tracks the probabilities on $O_i(t)$, conditional on its parents, and ignores other factors of the likelihood) is given by:

$$P^n(O_1, \ldots, O_n) \;\; = \;\; \prod_{i=1}^{n} \prod_{t=0}^{\tau+1} P_{L(t)}(L_i(t) \mid c_t^L(Pa(L_i(t))))g_t(A_it) \mid c_t^A(Pa(A_i(t)))).$$

Note that, even though subject i's data might be affected by data of other members of the target population of N subjects that are not included in the sample, the (i,t)-specific factor on the right-hand side is indeed a function of the observed data $O_i(t)$, and, in particular, the i-specific factor involving the product over time t is only a function of the observed data O_i, $i = 1, \ldots, n$. Recall that the likelihood for the full N individuals was given by

$$P(O) \;\; = \;\; \prod_{i=1}^{N} \prod_{t=0}^{\tau+1} P_{L(t)}(L_i(t) \mid c_t^L(Pa(L_i(t))))g_t(A_it) \mid c_t^A(Pa(A_i(t)))).$$

Thus, the partial likelihood corresponds with a random selection of factors of the actual likelihood of the full network of N individuals. It follows that the log-partial likelihood yields a valid loss function for the same conditional distributions of the full likelihood of the N individuals if the sample of n individuals is a true random sample of the set of N individuals. This proves that Q_0 can be expressed as a function of this partial likelihood P_0^n, showing the identifiability of Q_0.

Statistical model for observed data distribution: Each of the conditional distributions $P_{L(t)}$, $P_{A(t)}$ are unspecified beyond that they only depend on parents through $c_t^L(Pa(L(t)))$, $c_t^A(Pa(A(t)))$, respectively. This defines now a statistical model $\mathcal{M}^n$ for the partial likelihood P^n of $(O_1, \ldots, O_n)$ obtained by varying the choice of these conditional distributions.

Identification of causal quantity and the statistical target parameter: The conditional distributions $P_{L(t)}$, and thereby $P_{g^*,L(t)}$ are identified by the observed data distribution through maximizing the partial likelihood as $n \rightarrow \infty$ (and also $N \rightarrow \infty$), while we already established that $\psi_0^F = EY_{g^*}^c$ can be represented as a $\Psi(Q_0)$. Since Q_0 is identified by the partial likelihood P_0^n, it follows that the counterfactual mean average outcome $\psi_0^F = EY_{g^*}^c$ of $Y^c = 1/N \sum_i Y_i$ under a stochastic intervention g^* is identified by the true partial likelihood P_0^n of the observed data $O = (O_1, \ldots, O_n)$ as $n \rightarrow \infty$. This allows us to define a statistical target parameter $\Psi : \mathcal{M}^n \rightarrow \mathbb{R}$ on all possible partial likelihoods P^n in the model $\mathcal{M}^n$, where $\Psi(P_0^n) = \psi_0^F$ under our causal model.

To summarize: We note that the distribution of the counterfactual L_{g^*} defined by (1) is a factor over N factors identified by the conditional distributions of $L(t)$ and g^*. This expresses $EY_{g^*}^c$ as a function of these conditional distributions Q_0. Since we expressed these conditional distributions as a function of P_0^n, by simply plugging in this expression for these conditional distributions in (1) we obtain a mapping from P_0^n into $\Psi(P_0^n)$ defining our statistical target parameter. This defines a statistical model for the partial likelihood and a statistical target parameter as a feature of this partial likelihood, and thereby the statistical estimation problem.

Targeted maximum partial likelihood estimation: Since maximum partial likelihood estimation is completely analogue to the maximum likelihood estimator based on the full data on all N individuals we suggest that by replacing sums over i in the efficient influence curve for the case $n = N$ by sums over the actual n observed individuals we obtain the analogue of the efficient influence curve for the statistical target parameter. By the same token, we can define a similar TMLE as well, which we now call a targeted minimum loss based estimator since we are not maximizing an actual likelihood during the targeting step.

13 Discussion

We formulated a general causal model for the longitudinal data generated by a finite population of connected units. This allows us to define counterfactuals indexed by interventions on the treatment nodes of the units, and corresponding causal contrasts. We established identifiability of the causal quantities from the data observed on the units when observing all units or a random sample of the units, assuming that the size of the population converges to infinity, under appropriate assumptions. Our causal assumptions implied conditional independence across units at time t, conditional on the past of all units, resulting in a factorized likelihood of the observed data (even though the observed data is generated by a single experiment, not by a repetition of independent experiments). To deal with the curse of dimensionality we assumed that a unit's dependence on the past of other units can be summarized by a finite dimensional measure, and that this dependence is described by a common function across the units. This describes now the statistical model for the data distribution and thereby the statistical estimation problem. We demonstrated that we can use cross-validation and super-learning to estimate the different factors of the likelihood. Given the statistical model and statistical target parameter that identifies the counterfactual mean under an intervention, we derived the efficient influence curve of the target parameter. We showed that this efficient influence curve characterizes the normal limit distribution of a maximum likelihood estimator, and thus still represents an optimal asymptotic variance among estimators of the target parameter. However, due to the curse of dimensionality, maximum likelihood estimators will be ill-defined for finite samples, and smoothing will be needed.

Such smoothed/regularized maximum likelihood estimators are not targeted and will thereby be overly bias w.r.t. the target parameter, and, as a consequence, generally not result in asymptotically normally distributed estimators of the statistical target parameter. Therefore, we formulated targeted maximum likelihood estimators of this estimand, and showed that the robustness of the efficient influence curve implies that the bias of the TMLE will be a second order term involving squared differences of two nuisance parameters. Subsequently, we focussed on defining and analyzing the TMLE of causal effects of an intervention on a single treatment node on a future outcome. In this special case we showed that the efficient influence curve is double robust w.r.t. two nuisance parameters, one involving the intervention mechanism, and the other involving the common conditional mean function for the outcome. We established a formal asymptotic normality theorem under the assumption that each unit is only connected to fewer than K other units for a universal K.

In future work it will be of interest to extend this last theorem to the case that a unit can depend on summary measures across a number of units that can converge to infinity with sample size. It will also be of interest to allow that K depends on N, and establish rates of convergence that are slower than $1/\sqrt{N}$ and establish corresponding (e.g., normal) limit distributions. The finite sample behavior of these estimators and confidence intervals will need to be evaluated through simulation studies.

Overall, we believe that the statistical study of these causal models for networks of units provides a fascinating and important area of future research, relying on deep advances in empirical process and statistical estimation theory, while raising new challenges. In the mean time, these advances will be needed to move forward statistical practice.

References

J. Andersen, D. Faries, and R. Tamura. A randomized play-the-winner design for multiarm clinical trials. *Communication in Statistical Theory*, 23:309–323, 1994.

Z.D. Bai, F. Hu, and W.F. Rosenberger. Asymptotic properties of adaptive designs for clinical trials with delayed response. *Annals of Statistics*, 30(1): 122–139, 2002.

H. Bang and J.M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

D.A. Barr. The effects of organizational structure on primary care outcomes under managed care. *Annals of Internal Medicine*, 122:353359, 1995.

L.F. Berkman, T. Glass, I. Brissette, and T.E. Seeman. From social integration to health: Durkheim in the new millennium. *Social Science & Medicine*, 51: 843857, 2000.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1997.

A. Biglan, K. Smolkowski, T. Duncan, and C. Black. A randomised controlled trial of a community intervention to prevent adolescent tobacco use. *Tobacco Control*, 9:2432, 2000.

A.S. Bryk and S.W. Raudenbush. *Hierarchical linear models*. Newbury Park: Sage, 1992.

D.T. Campbell and J.C. Stanley. *Experimental and quasi-experimental designs for research*. Hopewell, NJ: Houghton Mifin, 1963.

M.J. Campbell, A. Donner, and N. Klar. Developments in cluster randomized trials and statistics in medicine. *Statistics in Medicine*, 26:2–19, doi: 10.1002/sim.2731, 2007a.

M.J. Campbell, A. Donner, and N. Klar. Developments in cluster randomized trials and statistics in medicine. *Statistics in Medicine*, 26:2–19, doi: 10.1002/sim.2731, 2007b.

J. Cassel. The contribution of the social environment to host resistance. *American Journal of Epidemiology*, 104:107–123, 1976.

A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate. Technical Report 258, Division of Biostatistics, University of California, Berkeley, 2010a.

A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, theoretical study. *Int J Biostat*, 7(1):1–32, 2011a. Working paper 258, www.bepress.com/ucbbiostat.

A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, simulation study. *Int J Biostat*, 7(1):33–, 2011b. Working paper 258,www.bepress.com/ucbbiostat.

A. Chambaz and M.J. van der Laan. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate. Technical Report 258, Division of Biostatistics, University of California, Berkeley, 2010b. Available at `http://biostats.bepress.com/ucbbiostat/paper258/`.

J.R. Charlton, M.F. DSouza, M. Tooley, and R. Silver. A community trial strategy for evaluating treatment for symptomatic conditions. *Statistics in Medicine*, 4:1121, 1985.

Y. Cheng and Y. Shen. Bayesian adaptive designs for clinical trials. *Biometrika*, 92(3):633–646, 2005.

J. Clark, D.A. Potter, and J.B. McKinlay. Bringing social structure back into clinical decision making. *Social Science & Medicine*, 32:853866, 1991.

C.C. Clogg and A. Haritou. The regression method for causal inference and a dilemma confronting this method. In *In S. P. Turner (Ed.), Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences*, page 83112. Notre Dame, IN: Notre Dame University Press, 1997.

W.G. Cochran. The planning of observational studies of human populations. *J R Stat Soc Ser A Gen*, 128(2):234–266, 1965.

J.B. Copas and H.G. Li. Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 59:5595, 1997.

A. Dawid and V. Didelez. Identifying the consequences of dynamic treatment strategies: A decision theoretic overview. *Statistics Surveys*, (4):184–231, 2010.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:18, 1977.

I. Diaz and M.J. van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics, to appear*, 2012.

V. Didelez, A.P. Dawid, and S. Geneletti. Direct and Indirect Effects of Sequential Treatments. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pages 138–146, Cambridge, MA, 2006.

A. Donner and N. Klar. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London, 2000.

D. Draper. Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20:115147, 1995.

H.A. Feldman, J. B. McKinlay, D.A. Potter, K.M. Freund, R.B. Burns, M.A. Moskowitz, and L.E. Kasten. Nonmedical inuences on medical decision making: An experimental technique using videotapes, factorial design, and survey sampling. *Health Services Research*, 32:343366, 1997.

H.A. Feldman, M.A. Proschan, D.M. Murray, D.C. Goff, M. Stylianou, E. Dulberg, P.G. McGovern, W. Chan, N.C. Mann, and V. Bittner. Statistical design of react (rapid early action for coronary treatment), a multisite community trial with continual data collection. *Controlled Clinicial Trials*, 19: 391403, 1998.

N. Flournoy and W.F. Rosenberger. *Adaptive Designs*. Hayward, Institute of Mathematical Statistics, 1995.

D. Freedman. From association to causation via regression (with comments). in s. p. turner (ed.). 1997.

R. Gill and J.M. Robins. Causal inference in complex longitudinal studies: continuous case. *Ann. Stat.*, 29(6), 2001.

H. Goldstein. *Multilevel statistical models in educational and social research.* London: Edward Arnold, 1987.

S. Greenland. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology*, 30:13431350, 2001.

S. Greenland. A review of multilevel theory for ecological analyses. *Statistics in Medicine*, 21:389395, 2002.

S. Greenland. Randomization, statistics, and causal inference. *Epidemiology*, 1:421429, 1990.

S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *International Journal of Biostatistics*, 6:article 26, www.bepress.com/ijb/vol6/iss1/26, 2010.

M.E. Halloran and C.J. Struchiner. Causal inference in infectious diseases. *Epidemiology*, 6:142151, 1995.

R.J. Hayes and L.H. Moulton. *Cluster Randomized Trials.* Chapman & Hall/CRC, Boca Raton, 2009.

J.J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47: 153162, 1979.

M.A. Hernán and J.M. Robins. *Causal Inference.* Chapman & Hall/CRC, 2012. Unpublished.

H.D. Holder, R.F. Saltz, A.J. Treno, J.W. Grube, and R.B. Voas. Evaluation design for a community prevention trial. an environmental approach to reduce alcohol- involved trauma. *Evaluation Review*, 21:140165, 1997.

P.W. Holland. Statistics and causal inference. *J Am Stat Assoc*, 81(396): 945–960, 1986.

E.B. Hook. (letter to editor) re: Neighborhood social environment and risk of death: Multilevel evidence from the alameda county study. *American Journal of Epidemiology*, 151:11321133, 2001.

F. Hu and W.F. Rosenberger. Analysis of time trends in adaptive designs withi appliation to a neurophysiology experiment. *Statistics in Medicine*, 19:2067–2075, 2000.

F. Hu and W.F. Rosenberger. *The theory of response adaptive randomization in clinical trials*. New York Wiley, 2006.

M.G. Hudgens and M.E. Halloran. Toward Causal Inference With Interference. *J Am Stat Assoc*, 103(482):832–842, 2008. PMCID: PMC2600548.

J. S. Kaufman and S. Kaufman. Assessment of structured socioeconomic effects on health. *Epidemiology*, 12:157167, 2001.

J.S. Kaufman and R.S. Cooper. Seeking causal explanations in social epidemiology. *American Journal of Epidemiology*, 150:113120, 1995.

J.S. Kaufman and C. Poole. Looking back on causal thinking in the health sciences. *Annual Review of Public Health*, 21:101119, 2000.

J.R. Koethe, A.O. Westfall, D.K. Luhanga, G.M. Clark, J.D. Goldman, P.L Mulenga, R.A. Cantrell, B.H. Chi, I. Zulu, M.S. Saag, and J. S. Stringer. A cluster randomized trial of routine hiv-1 viral load monitoring in zambia. *PlLoS One*, 5:5(3):e9680, 2010.

I.G.G. Kreft, J. de Leeuw, and R. van der Leeden. Review of five multilevel analysis programs: Bmdp-5, genmod, hlm, ml3, varcl. *American Statistician*, 48:324335, 1994.

N. Krieger. Theories for social epidemiology in the 21st century: An ecosocial perspective. *International Journal of Epidemiology*, 30:668677, 2001.

N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963974, 1982.

E. Leamer. Lets take the con out of econometrics. *American Economic Review*, 73:3243, 1983.

J. De Leeuw and I.G.G. Kreft. Software for multilevel analysis. In *In H. Goldstein (Ed.), Multilevel modelling of health statistics*, page 187204. New York: Wiley, 2001.

S.M. LeFort, K. Gray-Donald, K.M. Rowat, and M.E. Jeans. Randomized controlled trial of a community-based psychoeducation program for the self-management of chronic pain. *Pain*, 74:297306, 1998.

D. Lindley and A. Smith. Bayes estimation for linear models. *Journal of the Royal Statistical Society, Series B*, 34:141, 1972.

R.V. Luepker, J.M. Raczynski, S. Osganian, R.J Goldberg, Jr. J.R. Finnegan, J.R. Hedges, Jr. D.C. Goff, M.S. Eisenberg, J.G. Zapka, H.A. Feldman, D.R. Labarthe, P.G. McGovern, C.E. Cornell, M.A. Proschan, and D.G. Simons-Morton. Effect of a community intervention on patient delay and emergency medical service use in acute coronary heart disease: The rapid early action for coronary treatment (react) trial. *JAMA*, 284:6067, 2000.

G. Maldonado and S. Greenland. Estimating causal effects. *International Journal of Epidemiology*, 31:422438, 2002.

C.F. Manski. Identication problems in the social sciences. in p. v. marsden (ed.). In *Sociological methodology*. San Francisco: Jossey-Banks, 1993.

W.M. Mason, G.Y. Wong, and B. Entwisle. Contextual analysis through the multilevel linear model. in s. leinhardt (ed.). In *Sociological methodology: 19831984*. San Francisco: Jossey-Bass, 1984.

J.B. McKinlay. Some contributions from the social system to gender inequalities in heart disease. *Journal of Health and Social Behavior*, 37:126, 1996.

S.M. McKinlay. The design and analysis of the observational studya review. *Journal of the American Statistical Association*, 70:503523, 1975.

A.J. McMichael. Prisoners of the proximate: Loosening the constraints on epidemiology in an age of change. *American Journal of Epidemiology*, 149: 887897, 1999.

M.E.Sobel. Causal inference in the social and behavioral sciences. In *In M. E. Sobel (Ed.), Handbook of statistical modeling for the social and behavioral sciences*, page 138. New York: Plenum., 1995.

L. Meyer, N. Job-Spira, J. Bouyer, E. Bouvet, and A. Spira. Prevention of sexually transmitted diseases: A randomised community trial. *Journal of Epidemiology and Community Health*, 45:152158, 1991.

H. Morgenstern. Ecologic studies in epidemiology: Concepts, principles, and methods. *Annual Review of Public Health*, 16:6181, 1995.

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

J.M. Oakes. The (mis)estimation of neigborhood effects: causal inference for a practicable social epidemiology. *Social Science and Medicine*, 58:1929–1952, 2004.

T. Parsons. *The social system.* New York: Free Press, 1951.

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

J. Pearl. *Causality: models, reasoning, and inference.* Cambridge, New York, 2nd edition, 2009.

V. Persky, L. Coover, E. Hernandez, A. Contreras, J. Slezak, J. Piorkowski, L. Curtis, M. Turyk, V. Ramakrishnan, and P. Scheff. Chicago community-based asthma intervention trial: Feasibility of delivering peer education in an inner-city population. *Chest*, 116:216S223S, 1999.

M.L. Petersen and M.J. van der Laan. A general roadmap for the estimation of causal effects. Unpublished, Division of Biostatistics, University of California, Berkeley, 2012.

E.C. Polley, Sherri Rose, and M.J. van der Laan. Super learning. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer, New York Dordrecht Heidelberg London, 2012.

J.M. Robins. Data, design, and background knowledge in etiologic inference. *Epidemiology*, 12:313320, 2001.

J.M. Robins. Addendum to: “A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect” [Math. Modelling **7** (1986), no. 9-12, 1393–1512; MR 87m:92078]. *Comput. Math. Appl.*, 14(9-12):923–945, 1987a. ISSN 0097-4943.

J.M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chron Dis (40, Supplement)*, 2:139s–161s, 1987b.

J.M. Robins. Causal inference from complex longitudinal data. In Editor M. Berkane, editor, *Latent Variable Modeling and Applications to Causality*, pages 69–117. Springer Verlag, New York, 1997.

J.M. Robins. [choice as an alternative to control in observational studies]: Comment. *Statistical Science*, 14(3):281–293, 1999.

J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*. Birkhäuser, Basel, 1992.

G. Rose. Sick individuals and sick populations. *International Journal of Epidemiology*, 14:3238, 1985.

P.R. Rosenbaum. *Observational studies*. Springer, Berlin Heidelberg New York, 2nd edition, 2002.

W.F. Rosenberger. New directions in adaptive designs. *Statistical Science*, 11: 137–149, 1996.

W.F. Rosenberger and S.E. Grill. A sequential design for psychophysical experiments: An application to estimating timing of sensory events. *Statistics in Medicine*, 16:2245–2260, 1997.

W.F. Rosenberger and T.N. Shiram. Estimation for an adaptive allocation design. *Journal of Statistical Planning and Inference*, 59:309–319, 1997.

W.F. Rosenberger, N. Flournoy, and S.D. Durham. Asymptotic normality of maximum likelihood estimators from multiparameter response driven designs. *Journal of Statistical Planning and Inference*, pages 69–76, 1997.

M. Rosenblum and M.J. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *Int J Biostat*, 6(2): Article 19, 2010.

D.B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge, MA, 2006.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ Psychol*, 64:688–701, 1974.

D.B. Rubin. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 47: 12131234, 1991.

R.H. Shipley, T.D. Hartwell, W.D. Austin, A.C. Clayton, and L.C. Stanley. Community stop-smoking contests in the commit trial: Relationship of participation to costs. *Community intervention trials. Preventive Medicine*, 24: 286292, 1995.

J.D. Singer. Using sas proc mixed to t multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 24:323355, 1998.

H.L. Smith. Specication problems in experimental and nonexperimental social research. In *In C. C. Clogg (Ed.), Sociological methodology*, volume 20. Oxford: Basil, 1990.

P. Starr. *The social transformation of American medicine.* New York: Basic Books, 1982.

M. Susser. *Causal thinking in health sciences: Concepts and strategies of epidemiology.* New York: Oxford, 1973.

M. Susser. Should the epidemiologist be a social scientist or a molecular biologist? *International Journal of Epidemiology*, 28:10191022, 1999.

R.N. Tamura, D.E. Faries, J.S. Andersen, and J.H. Heiligenstein. A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *Journal of the American Statistical Association*, 89:768–776, 1994.

E.J. Tchetgen Tchetgen and T.J. VanderWeele. On causal inference in the presence of interference. *Stat Meth Med Res*, 21(1):55–75, 2012. PMID: 21068053.

M. Toftager, L.B. Christiansen, P.L. Kristensen, and J. Troelsen. Space for physical activity-a multicomponent intervention study: study design and baseline findings from a cluster randomized controlled trial. *BMC Public Health*, 10:711–777, 2011.

A.A. Tsiatis. Information based monitoring of clinical trials. *Statistics in Medicine*, 2006.

M.J. van der Laan. The construction and analysis of adaptive group sequential designs. Technical report 232, Division of Biostatistics, University of California, Berkeley, March 2008.

M.J. van der Laan. Targeted maximum likelihood based causal inference: Part I. *Int J Biostat*, 6(2):Article 2, 2010a.

M.J. van der Laan. Targeted maximum likelihood based causal inference: Part II. *Int J Biostat*, 6(2):Article 3, 2010b.

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, 2003.

M.J. van der Laan and S. Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int J Biostat*, 8(1), 2012. PMID: 22611591.

M.J. van der Laan and M.L. Petersen. Randomized trials and observational studies in which treatment allocation depends on the baseline characteristics of all sampled units. Technical Report 296, Division of Biostatistics, University of California, Berkeley, 2012. Available at `http://biostats.bepress.com/ucbbiostat/paper296/`.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, Berlin Heidelberg New York, 2003.

M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2012.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.

M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

A.W. van der Vaart. *Asymptotic statistics*. Cambridge, New York, 1998.

A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3):351–371, 2006.

T.J. VanderWeele, J.P. Vandenbrouke, E.J. Tchetgen Tchetgen, and J.M. Robins. A mapping between interactions and interference: implications for vaccine trials. *Epidemiology*, 23(3):285–292, 2012. PMID: 22317812.

G. Verbeke and G. Molenbergs. *Linear mixed models in practice: A SAS oriented approach.* New York: Springer, 1997.

L. Watson, R. Small, S. Brown, W. Dawson, and J. Lumley. Mounting a community-randomized trial: sample size, matching, selection, and randomization issues in prism. *Journal of Controlled Clinical Trials*, 3:235–250, 2004.

L.J. Wei. The generalized polya's urn design for sequential medical trials. *Annals of Statistics*, 7:291–296, 1979.

L.J. Wei and S. Durham. The randomize play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73:840–843, 1978.

L.J. Wei, R.T. Smythe, D.Y. Lin, and T.S. Park. Statistical inference with data-dependent treatment allocation rules. *Journal of the American Statistical Association*, 85:156–162, 1990.

C. Winship and S.L. Morgan. The estimation of causal effects from observational data. *Annu Rev Sociol*, 25:659–707, 1999.

Z. Yu and M.J. van der Laan. Double robust estimation in longitudinal marginal structural models. Technical report, Division of Biostatistics, University of California, Berkeley, 2003.

M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64:131–146, 1969.

W. Zheng and M.J. van der Laan. Causal mediation in a survival setting with time-dependent mediators. Technical Report 295, Division of Biostatistics, University of California, Berkeley, 2012. Available at `http://biostats.bepress.com/ucbbiostat/paper295/`.

Appendix

Introduction to Appendix

We start out with presenting a general template of our proof of Theorem 4 which establishes the asymptotics of the TMLE for the case $\tau = 0$. In this template we define the remaining ingredients (A1), (A2) and (A3) that will need to be established in the remainder of the proof. Each of these three ingredients is carried out in a separate section. These sections are themselves organized by special tasks that need to be carried out. We conclude with demonstrating how our template can be generalized to the analysis of the TMLE for general longitudinal data structures (i.e., arbitrary integer values τ).

A General template of proof of Theorem 4.

Recall that $D^* = 1/N \sum_{j=1}^N D_j^*(O)$ is a sum over the individuals j. We will use the notation $P_N D^* = 1/N \sum_{j=1}^N D_j^*(O)$, while $P_0 D^* = 1/N \sum_j E_{P_0} D_j^*(O)$ is its expectation w.r.t. distribution P_0. We have $D^* = D_W^* + D_Y^*$, $P_0 D_W^*(\bar{Q}_N^*, Q_{W,0}) = 0$, $P_0 D_Y^*(\bar{Q}_N^*, Q_{W,0}) = \psi_0 - \Psi(\bar{Q}_N^*, Q_{W,0})$, and $P_N D_W^*(\bar{Q}_N^*, Q_{W,N}) = P_N D_Y^*(\bar{Q}_N^*, Q_{W,N}) = 0$. In particular, this yields

$$P_0 D^*(Q_{Y,N}^*, Q_{W,0}) = \psi_0 - \Psi(\bar{Q}_N^*, Q_{W,0}).$$

We now proceed as follows:

$$\begin{aligned}
&\Psi(\bar{Q}_N^*, Q_{W,N}) - \psi_0 = \Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) + \Psi(\bar{Q}_N^*, Q_{W,0}) - \psi_0 \\
&= \Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) - P_0 D_Y^*(\bar{Q}_N^*, Q_{W,0}) \\
&= \Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) + (P_N - P_0) D_Y^*(\bar{Q}_N^*, Q_{W,0}) \\
&\qquad\qquad - P_N\{D_Y^*(\bar{Q}_N^*, Q_{W,0}) - D_Y^*(\bar{Q}_N^*, Q_{W,N})\} \\
&= \Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) + (P_N - P_0) D_Y^*(\bar{Q}_N^*, Q_{W,0}) \\
&\qquad\qquad + (P_N - P_0)\{D_Y^*(\bar{Q}_N^*, Q_{W,N}) - D_Y^*(\bar{Q}_N^*, Q_{W,0})\} \\
&\qquad\qquad + P_0\{D_Y^*(\bar{Q}_N^*, Q_{W,N}) - D_Y^*(\bar{Q}_N^*, Q_{W,0})\}.
\end{aligned}$$

We note that

$$\{D_Y^*(\bar{Q}_N^*, Q_{W,N}) - D_Y^*(\bar{Q}_N^*, Q_{W,0})\} = \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(Y - \bar{Q}_N^*),$$

where $\bar{h}_N^* = \bar{h}(g^*, Q_{W,N})$, $\bar{h}_0^* = \bar{h}(g^*, Q_{W,0})$, and similarly for $\bar{h}_0^*$, $\bar{h}_0$. We also note that

$$\begin{aligned}
&\Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) \\
&= \tfrac{1}{N} \textstyle\sum_{i=1}^N \left\{\int_a \bar{Q}_N^*(c_i^Y(a, W)) g^*(a \mid W) - \int \bar{Q}_N^*(c_i^Y(a, w)) g^*(a \mid w) Q_{W,0}(w)\right\} \\
&\equiv \tfrac{1}{N} \textstyle\sum_{i=1}^N \{f_{W,i}^1(W) - P_0 f_{W,i}^1\} + R_{N,0},
\end{aligned}$$

where

$$f_{W,i}^1 = \int_a \bar{Q}_Y^*(c_i^Y(a, W)) g^*(a \mid W),$$

and

$$R_{N,0} = \frac{1}{N}\sum_{i=1}^{N} \left\{ \int_a (\bar{Q}_N^* - \bar{Q}^*)(c) g_i^*(c \mid W) - \int (\bar{Q}_N^* - \bar{Q}^*)(c) g_i^*(c \mid w) Q_{W,0}(w) \right\}.$$

We used here that $\int_a \bar{Q}(c_i^Y(a,W)) g^*(a \mid W) = \int_c \bar{Q}(c) g_i^*(c \mid W)$. Define the process $Z^1_{W,N}(\bar{Q}) = \frac{1}{N}\sum_{i=1}^N \{ f^1_{W,i}(\bar{Q}) - P_0 f^1_{W,i}(\bar{Q})\}$ indexed by $\bar{Q}$, where $f^1_{W,i}(\bar{Q}) = \int \bar{Q}(c) g_i^*(c \mid W)$. Note that $R_{N,0} = Z^1_{W,N}(\bar{Q}_N^* - \bar{Q}^*)$. As a consequence, showing that $R_{N,0} = o_P(1/\sqrt{N})$ corresponds with proving that $Z^1_{W,N}(\epsilon_N) = o_P(1/\sqrt{N})$ for a sequence ϵ_N that converges to zero w.r.t. some norm. Therefore, our proof will involve studying this process $Z^1_{W,N}$ and establishing the required asymptotic equicontinuity. In this manner, we will establish

$$R_{N,0} = o_P(1/\sqrt{N})(A2).$$

Thus, we have obtained the following expansion:

$$\begin{aligned} \psi_N^* - \psi_0 = \tfrac{1}{N}\textstyle\sum_{i=1}^N \{ f^1_{W,i}(W) - P_0 f_{W,i}\} + (P_N - P_0) D_Y^*(\bar{Q}_N^*, Q_{W,0}) \\ + P_0 \left(\tfrac{\bar{h}_N^*}{\bar{h}_N} - \tfrac{\bar{h}_0^*}{\bar{h}_0} \right) (\bar{Q}_{Y,0} - \bar{Q}_Y^*) + R_N, \end{aligned}$$

where

$$\begin{aligned} R_N &= -P_0 \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0} \right) (\bar{Q}_N^* - \bar{Q}_Y^*) \\ &\quad +(P_N - P_0) \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0} \right) (Y - \bar{Q}_N^*) \\ &\equiv R_{N,1} + R_{N,2}. \end{aligned}$$

We assumed that the second order term $R_{N,1} = o_P(1/\sqrt{N})$.

We have

$$\begin{aligned} &(P_N - P_0) D_Y^*(\bar{Q}_N^*, Q_{W,0}) = (P_N - P_0) D_Y^*(\bar{Q}^*, Q_{W,0}) \\ &+(P_N - P_0)\{ D_Y^*(\bar{Q}_N^*, Q_{W,0}) - D_Y^*(\bar{Q}^*, Q_{W,0})\} \\ &\equiv (P_N - P_0) D_Y^*(\bar{Q}^*, Q_{W,0}) + R_{N,3}. \end{aligned}$$

We will show that

$$R_{N,2} = o_P(1/\sqrt{N}) \text{ and } R_{N,3} = o_P(1/\sqrt{N})(A3).$$

Define the process $Z_N(\tilde{h}, \bar{Q}) = (P_N - P_0)\tilde{h}(Y - \bar{Q})$, which is a sum of the form $Z_N(\tilde{h}, \bar{Q}) = \sum_i f_i(\bar{Q}, h)(O_i)$ indexed by $(\tilde{h}, \bar{Q})$, where $\tilde{h}$ plays role of $\bar{h}^*/\bar{h}$. Showing that $R_{N,3} = o_P(1/\sqrt{N})$ and that $R_{N,2} = o_P(1/\sqrt{N})$ requires showing that $Z_N(\epsilon_N) = o_P(1/\sqrt{N})$ for ϵ_N converging to zero w.r.t some norm. Therefore our proof will involve studying this process, and establishing the required asymptotic equicontinuity. Specifically, we will decompose this process in three orthogonal processes that can be represented as sums over functions of conditionally independent random variables identified by the sets F_i (analogue to orthogonal decomposition below of the first order approximation), and establish this asymptotic equicontinuity for each of the three orthogonal processes.

Consider now the term

$$P_0 \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0} \right) (\bar{Q}_{Y,0} - \bar{Q}_Y^*). \tag{7}$$

We have

$$\begin{aligned}
&P_0 \tfrac{1}{N} \textstyle\sum_{i=1}^N \left\{ \frac{\bar{h}(g^*,Q_{W,N})(c_i^Y)}{\bar{h}(g_0,Q_{W,N})(c_i^Y)} - \frac{\bar{h}(g^*,Q_{W,0})(c_i^Y)}{\bar{h}(g_0,Q_{W,0})(c_i^Y)} \right\} (\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c_i^Y) \\
&= \textstyle\int_c \left\{ \frac{\bar{h}(g^*,Q_{W,N})(c)}{\bar{h}(g_0,Q_{W,N})(c)} - \frac{\bar{h}(g^*,Q_{W,0})(c)}{\bar{h}(g_0,Q_{W,0})(c)} \right\} (\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c)\bar{h}_0(c) \\
&= \textstyle\int_c \left\{ \frac{\bar{h}_N^* - \bar{h}_0^*}{\bar{h}_0}(c) - \frac{\bar{h}_0^*}{\bar{h}_0^2}(\bar{h}_N - \bar{h}_0)(c) \right\} (\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c)\bar{h}_0(c) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + R_{N,4},
\end{aligned}$$

where

$$R_{N,4} = \int_c \left\{ \frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}} \right\} \frac{1}{\bar{h}_0} (\bar{h}_N - \bar{h}_0)(\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c)\bar{h}_0(c).$$

We assumed that $R_{N,4} = o_P(1/\sqrt{N})$. We now note that

$$\begin{aligned}
&\bar{h}_0(c) = \textstyle\sum_{i=1}^N P(c_i^Y(A,W) = c) \\
&= \textstyle\sum_{i=1}^N \int I(c_i^Y(a,w) = c) g_0(a \mid w) Q_{W,0}(w) \\
&= \textstyle\sum_{i=1}^N h_i(g_0, Q_{W,0})(c) \\
&\bar{h}_N(c) = \textstyle\sum_{i=1}^N \int I(c_i^Y(a,w) = c) g_0(a \mid w) Q_{W,N}(w) \\
&= \textstyle\sum_{i=1}^N \int I(c_i^Y(a,W) = c) g_0(a \mid W) \\
&= \textstyle\sum_{i=1}^N h_i(g_0, Q_{W,N})(c)
\end{aligned}$$

Thus, we can conclude that (7) reduces to

$$\begin{aligned}
&\tfrac{1}{N} \textstyle\sum_{i=1}^N \int_c \left\{ \frac{h_i(g^*,Q_{W,N})}{\bar{h}_0}(c) - \frac{\bar{h}_0^*}{\bar{h}_0^2} h_i(g_0, Q_{W,N}) \right\} (\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c)\bar{h}_0(c) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + o_P(1/\sqrt{N}) \\
&\equiv \tfrac{1}{N} \textstyle\sum_{i=1}^N f_{W,i}^2(W) + o_P(1/\sqrt{N}) \\
&\equiv Z_{W,N}^2 + o_P(1/\sqrt{N}),
\end{aligned}$$

where we note that $P_0 f^2_{W,i}(W) = 0$. The term $Z^2_{W,N} = 1/N \sum_i \{f^2_{W,i}(W) - P_0 f^2_{W,i}(W)\}$ is included in the first order expansion and thus partly characterizes the normal limit distribution of ψ^*_N, so that its analysis will be part of the analysis of the first order approximation.

Since $h_i(g, Q_{W,N})$ only depends on $(W_1, \ldots, W_N)$ through $(W_j : j \in F_i)$, where we condition on $F_1, \ldots, F_N$, we will indeed be able to show that a term $Z^2_{W,N}$ is itself a normally distributed random variable, even though each i-specific term is correlated with the j-specific terms when $F_i \cap F_j \neq \emptyset$.

Thus, if we prove (A2) and (A3), then we have obtained the following first order expansion:

$$\begin{aligned}\psi^*_N - \psi_0 &= \frac{1}{N}\sum_{i=1}^{N}\{f^1_{W,i}(W) - P_0 f^1_{W,i}\} + \frac{1}{N}\sum_i \{f^2_{W,i}(W) - P_0 f^2_{W,i}\} \\ &\quad +(P_N - P_0) D^*_Y(\bar{Q}^*_Y, Q_{W,0}) + o_P(1/\sqrt{N}).\end{aligned}$$

Analysis of first order approximation: Let $\bar{f}_{W,i} = f^1_{W,i} + f^2_{W,i}$. The first order approximation equals

$$\begin{aligned}&1/N \textstyle\sum_i \{D^*_{Y,i}(\bar{Q}^*, Q_{W,0}) + \bar{f}_{W,i}(W) - P_0\{D^*_{Y,i} + \bar{f}_{W,i}\}\} \\ &\equiv 1/N \textstyle\sum_i f_i(O).\end{aligned}$$

Using our notation, this thus equals $(P_N - P_0)f$. It remains to prove that this first order expansion converges to a normal limit distribution. This proof has its own outline. Firstly, we decompose $1/N \sum_i f_i(O)$ by $f_i = f_{i,W} + f_{i,A} + f_{i,Y}$, where $f_{i,W} = E_0(f_i \mid W) - E_0 f_i$, $f_{i,W} = E_0(f_i \mid W) - E_0 f_i$, $f_{i,A} = E_0(f_i \mid A, W) - E_0(f_i \mid W)$, and $f_{i,Y} = f_i - E_0(f_i \mid A, W)$. Denote the three corresponding terms with $Z_{Y,N} + Z_{A,N} + Z_{W,N}$. Note that $Z_{W,N} = 1/N \sum_i \{f_{W,i} - P_0 f_{W,i}\}$, where $f_{W,i} = \bar{f}_{W,i} + E(D^*_{Y,i} \mid W)$. It follows that $f_{W,i}$ simplifies to:

$$\begin{aligned}f_{W,i}(W) - P_0 f_{W,i} &= \int_a \bar{Q}_{Y,0}(c^Y_i(a, W)) g^*(a \mid W) \\ &\quad - \int_{a,w} \bar{Q}_{Y,0}(c^Y_i(a, W)) g^*(a, W) Q_{W,0}(w) \\ &= \int_c \bar{Q}_{Y,0}(c) g^*_i(c \mid W) - \int_{c,w} \bar{Q}_{Y,0}(c) g^*_i(c \mid W) Q_{W,0}(w).\end{aligned}$$

We also note that, conditional on W, A, Z^N_Y is a sum of independent mean zero random variables $f_{Y,i}(Y_i)$ (functions of Y_i), conditional on W, $Z^N_A = 1/\sqrt{N} \sum_i f_{A,i}((A_j : j \in F_i)) - P_0 f_{A,i}$ for some $f_{A,i}$, where A_i, $i = 1, \ldots, N$

are pairwise (conditionally) independent, and $Z_W^N = 1/\sqrt{N}\sum_i f_{W,i}(W_j : j \in F_i) - P_0 f_{W,i}$, W_i are pairwise independent.

We will show that

$$\begin{array}{l} Z_Y^N \Rightarrow_d N(0,\sigma_Y^2) \\ Z_A^N \Rightarrow_d N(0,\sigma_A^2) \\ Z_W^N \Rightarrow_d N(0,\sigma_W^2) \text{ (A1)} \end{array}$$

with the expressions for σ_Y^2, σ_A^2 and σ_W^2 as specified in the Theorem. Here (A1) represents all three convergence statements. Due to the orthogonality of the three empirical processes, using moment generating functions, it also follows that $Z_Y^N + Z_A^N + Z_W^N \Rightarrow_d N(0, \sigma^2 = \sigma_Y^2 + \sigma_A^2 + \sigma_W^2)$. For example, we can analyze $E(Z_Y^N + Z_A^N)^p = \sum_t c(p,k) E\{Z_Y^N\}^k E\{Z_A^N\}^{p-k}$ and use convergence of moments of each process separately to establish convergence to $E(Z_Y + Z_A)^p$. Once we have convergence of all moments, and we can bound $E(Z_Y^N + Z_A^N)^p \leq C^p$ for some $C < \infty$, which follows from our separate analysis, then we obtain convergence in moment generating function, and thereby weak convergence.

This finishes the outline of the proof. It remains to establish (A1), (A2), (A3).

B (A3)

B.1 (A3): Outline of Proof.

Let $\tilde{h} = \bar{h}^*/\bar{h}$ and we will denote D_Y^* with $D_Y^*(\tilde{h}, \bar{Q})$. Our goal is to prove that $\sqrt{N}(P_N - P_0)\{D_Y^*(\tilde{h}_0, \bar{Q}_N^*) - D^*(\tilde{h}_0, \bar{Q}^*)\} = o_P(1)$ and $\sqrt{N}(P_N - P_0)\{D_Y^*(\tilde{h}_N, \bar{Q}_N^*) - D_Y^*(\tilde{h}_0, \bar{Q}_N^*)\} = o_P(1)$. Let $P_{0,Y|A,W}f$, $P_{0,A|W}f$, $P_{0,W}f$, denote the expectation operators w.r.t their respective conditional distributions. We have

$$\begin{array}{l} Z_N(\tilde{h},\bar{Q}) = \sqrt{N}(P_N - P_0)D_Y^*(\tilde{h},\bar{Q}) = \sqrt{N}(P_N - P_{0,Y|A,W})D_Y^*(\tilde{h},\bar{Q}) \\ +\sqrt{N}(P_N - P_{0,A|W})P_{0,Y|A,W}D_Y^*(\tilde{h},\bar{Q}) \\ +\sqrt{N}(P_N - P_{0,W})P_{0,A|W}P_{0,Y|A,W}D_Y^*(\tilde{h},\bar{Q}) \\ = \frac{1}{\sqrt{N}}\sum_{i=1}^N \tilde{h}(c_i^Y)(Y_i - \bar{Q}_{Y,0}(c_i^Y)) \\ +\frac{1}{\sqrt{N}}\sum_{i=1}^N \tilde{h}(c_i^Y)(\bar{Q}_{Y,0} - \bar{Q})(c_i^Y) - \int_c \tilde{h}(c)(\bar{Q}_{Y,0} - \bar{Q})(c) g_{0,i}(c \mid W) \\ +\frac{1}{\sqrt{N}}\sum_{i=1}^N \int_c \tilde{h}(\bar{Q}_{Y,0} - \bar{Q})(c) g_{0,i}(c \mid W) - P_0 D_Y^*(\tilde{h},\bar{Q}) \\ \equiv Z_Y^N(\tilde{h}) + Z_A^N(\tilde{h},\bar{Q}) + Z_W^N(\tilde{h},\bar{Q}). \end{array}$$

We now note that, for a fixed $(\tilde{h},\bar{Q})$, conditional on W, A, Z_Y^N is a sum of independent mean zero random variables $f_{Y,i}(Y_i)$ (functions of Y_i). We also note that for a fixed $(\tilde{h},\bar{Q})$, conditional on W, $Z_A^N = 1/\sqrt{N}\sum_i f_{A,i}((A_j : j \in F_i)) - P_0 f_{A,i}$ for some $f_{A,i}$, where A_i, $i = 1, \ldots, N$ are pairwise (conditionally) independent. Finally, for a fixed $(\tilde{h}, \bar{Q})$, $Z_W^N = 1/\sqrt{N} \sum_i f_{W,i}(W_j : j \in F_i) - P_0 f_{W,i}$, W_i are pairwise independent.

Let $\bar{Q}^*$ be the limit of $\bar{Q}_N$, and let $\tilde{h}_0 = \bar{h}_0^* / \bar{h}_0$ be the limit of $\tilde{h}_N$. We will use empirical process theory to establish that

$$\begin{aligned} Z_Y^N(\tilde{h}_N) &= Z_N^Y(\tilde{h}_0) + o_P(1) \\ Z_A^N(\tilde{h}_N, \bar{Q}_N^*) &= Z_A^N(\tilde{h}_0, \bar{Q}^*) + o_P(1) \\ Z_W^N(\tilde{h}_N, \bar{Q}_N^*) &= Z_W^N(\tilde{h}_0, \bar{Q}^*) + o_P(1). \end{aligned}$$

This then establishes $R_{N,2} = o_P(1/\sqrt{N})$ and $R_{N,3} = o_P(1/\sqrt{N})$.

B.2 (A3): Outline of establishing asymptotic equicontinuity of a process

For that purpose, we will apply Lemma 5 in van der Vaart and Wellner (1996), which concerns establishing weak convergence of a process $(Z^N(f) : f \in \mathcal{F})$, indexed by a $f = (\tilde{h}, \bar{Q}) \in \mathcal{F}$. Given that $\mathcal{F}$ is a subset of some metric space of functions with metric d, one defines $N(\epsilon, \mathcal{F}, d)$ as the minimal number of balls of size ϵ needed to cover $\mathcal{F}$. In addition, let $\parallel Z^N(f) \parallel_\psi = \inf\{c_0 : \psi(\mid Z^N(f) \mid / c_0) \leq 1\}$ be the orlics norm of the random variable $Z^N(f)$.

For example, one can select the L^p-norm $\parallel Z^N(f) \parallel_p = \left\{\int E\{Z^N(f)\}^p\right\}^{1/p}$of $Z^N(f)$ for arbitrary large p which correspond with the choice of orlics norm defined by $\psi_p(x) = x^p$. The orlics norm implied by $\psi_{2,e}(x) = \exp(x^2) - 1$ is the typical orlics norm pursued in the case of sums of independent random variables, and this is the one we will also use.

This Lemma 5 states that, if 1) $\parallel Z^N(f_1) - Z^N(f_2) \parallel_\psi$ is bounded by $cd(f_1, f_2)$ for some universal constant c and metric $d(f_1, f_2)$, 2) $\mathcal{F}$ is totally bounded w.r.t.. this metric d, 3) for some $\eta > 0$, $\int_0^\eta \psi^{-1}\left(N(\epsilon, \mathcal{F}, d)\right) d\epsilon < \infty$, 4) the marginal distributions $Z^N(f)$ converge to a normal distribution $Z(f)$, then Z^N converges weakly to a Gaussian process Z in $\ell^\infty(\mathcal{F})$. We assumed that our parameter space $\mathcal{F}$ for $(\tilde{h}_0, \bar{Q}_{Y,0})$ and its difference $\mathcal{F}_d = \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}\}$ consists of uniformly bounded functions on a set $\mathcal{C}^Y$ that contains $C_i^Y(A, W)$ with probability 1, so that 2) holds. We posed 3) as an entropy condition on the parameter space $\mathcal{F}$, which will thus hold by assumption. For example, $\mathcal{F}$ could be the class of functions on $\mathcal{C}^Y \subset \mathbb{R}^d$ that have uniform sectional variation norm bounded by a $M < \infty$, in which case this entropy condition holds. Under conditons 1-3 we have that the process Z^N is asymptotically tight, and, for any sequence $\delta_n \to 0$, we have for each $x > 0$,

$$P\left(\sup_{d(f_1,f_2)<\delta_n} \mid Z^N(f_1) - Z^N(f_2) \mid > x\right) \to 0 \text{ as } N \to \infty.$$

So once we have established the orlics-norm condition 1), then this tightness can be used to establish that terms $Z^N(f_N) - Z^N(f) = o_P(1)$ for random f_N converging to f w.r.t. metric d in probability.

B.3 Bounding the orlics norm of our empirical processes.

The orlics norm $\| \cdot \|_\psi$ indexed by function $\psi(x) = \exp(x) - 1$ is defined as

$$\| X \|_\psi = \inf \{c > 0 : E \exp(\mid X \mid /C) - 1 \leq 1\} .$$

Similarly, we define the orlics norm for function $\psi_2(x) = \exp(x^2) - 1$. We consider a stochastic process $X_N(f)$ indexed by $f \in \mathcal{F}$ for a class of functions $\mathcal{F}$. In our application we have that $f = (\bar{Q}, \tilde{h}) \in \mathcal{F}$ represents two real valued functions $\bar{Q}$ and $\tilde{h}$ defined on a support $\mathcal{C}^Y \subset \mathbb{R}^d$ of $c^Y(A, W)$. In addition, our processes can be represented as $X_n(f) = 1/\sqrt{N} \sum_{i=1}^N f_i(f)(O)$, where for each i there is an associated set $F_i \subset \{1, \ldots, N\}$, and, if $F_i \cap F_j = \emptyset$, then $f_i(f)(O)$ and $f_j(f)(O)$ are independent. For some of our processes, these independencies are conditional on some random variables. Therefore, we would need to apply our general proof below conditional on these random variables, and then take the expectation of the resulting bound, still providing us with the desired marginal bound on the orlics norm. Thus, we need to bound $\| X_N(f) \|_\psi \leq C \| f \|$ for some universal (in N and $f \in \mathcal{F}$) $C < \infty$. As outlined in previous subsection, the choice of orlics norm and norm $\| f \|$ for $f \in \mathcal{F}$ is important, since the corresponding entropy requirement on $\mathcal{F}$ is that $\int_0^\eta \psi^{-1}(N(\epsilon, \| \cdot \|, \mathcal{F}))d\epsilon < \infty$. We will establish our results for the strongest orlics norm which corresponds with $\psi_2(x)$, while we select the supremum norm $\| f \| = \max(\| \bar{Q} \|_\infty, \| \tilde{h} \|_\infty)$ for the functions f.

Lemma 2 *Let $\| X \|_\psi$ be the orlics norm defined above w.r.t. $\psi(x) = \exp(x^2) - 1$. Suppose that for each p*

$$E \mid X_N(f) \mid^p \leq C(N, p) \| f \|^p .$$

Let $D(N)$ be a number so that

$$\sum_{p=1}^\infty C(N, 2p) D(N)^{2p}/p! \leq 1.$$

Then,

$$\| X_n(f) \|_\psi \leq \frac{1}{D(N)} \| f \| .$$

In particular, if $C(N,p)$ can be bounded from above by $C(p)$ constant in N, and one finds a D (constant in N) so that $\sum_{p=1}^{\infty} C(2p)D^{2p}/p! \leq 1$, then follows that $\parallel X_n(f) \parallel_{\psi} \leq \frac{1}{D} \parallel f \parallel$.

Proof. We first note

$$E \exp\{(X_N(f)/C)\}^2 - 1 = \sum_{p=1}^{\infty} \frac{E(|X_N(f)|/C)^{2p}}{p!} = \sum_{p=1}^{\infty} \frac{E|X_N(f)|^{2p}}{C^{2p}p!}.$$

Suppose that for each even p $E \mid X_N(f) \mid^p \leq C(N,p) \parallel f \parallel^p$. Then, we have

$$E \exp(X_N(f)/C) - 1 \leq \sum_{p=1}^{\infty} \frac{C(N,2p) \parallel f \parallel^{2p}}{C^{2p}p!}.$$

So $\parallel X_n(f) \parallel_{\psi}$ is bounded by a C so that

$$\sum_{p=1}^{\infty} \frac{C(N,2p)}{p!} \left(\frac{\parallel f \parallel}{C} \right)^{2p} \leq 1.$$

Let $D(N)$ be a number so that

$$\sum_{p=1}^{\infty} C(N,2p)D(N)^{2p}/p! \leq 1.$$

Then, C can be selected so that $\parallel f \parallel /C \leq D(N)$, or equivalently, $C \geq \parallel f \parallel /D(N)$. Thus, we have shown that $\parallel X_N(f) \parallel_{\psi} \leq \frac{1}{D(N)} \parallel f \parallel$. The last statement is straightforwardly shown. This completes the proof. □

Thus, apparently, it suffices to establish a bound of the type $E \mid X_N(f) \mid^p \leq C(p) \parallel f \parallel^p$ for some $C(p)$ that is somewhat well behaved as function in p for $p \rightarrow \infty$ so that the previous lemma applies.

We use the following lemma to bound the p-th moment of $X_N(f)$.

Lemma 3 *Assume that, for each $i = 1, \ldots, N$, and each integer p, we have a universal constant C so that*

$$\parallel f_i(f) \parallel_p = (E(f_i(f)(O))^p)^{1/p} \leq C \parallel f \parallel . \tag{8}$$

Then, we have

$$E \prod_{j=1}^{p} f_j \leq \prod_{j=1}^{p} \parallel f_j \parallel_{2j} \leq C^p \parallel f \parallel^p .$$

The bounding (8) is a straightforward consequence of our conditions stated in the theorem.

Proof. By repeatedly applying Cauchy-Schwarz inequality, it follows that

$$E\prod_j f_j(f) \leq \prod_{j=1}^p \left(Ef_j(f)^{2j}\right)^{1/(2j)} = \prod_{j=1}^p \parallel f_j(f) \parallel_{2j} .$$

By assumption, $\parallel f_j(f) \parallel_{2j} \leq C \parallel f \parallel$, so that the latter is bounded by $C^p \parallel f \parallel^p$. □

The following lemma provides us with an upper bound for $C(N,p)$ so that $E \mid X_N(f) \mid^p \leq C(N,p) \parallel f \parallel$.

Lemma 4 *Assume that, for each i, and each integer p, we have a universal constant C so that*

$$\parallel f_i(f) \parallel_p = (E(f_i(f)(O))^p)^{1/p} \leq C \parallel f \parallel .$$

Given an index $\vec{i} = (i_1, \ldots, i_p) \in \{1, \ldots, N\}^p$ (one among N^p elements), we can draw a graph by drawing a line between any two elements i_{l_1}, i_{l_2} in $\{i_1, \ldots, i_p\}$ whenever the two corresponding sets $F(i_{l_1})$ and $F(i_{l_2})$ have a non-empty intersection. Let $R(i_1, \ldots, i_p)$ be an indicator, identified by indices $\vec{i} = (i_1, \ldots, i_p) \in \{1, \ldots, N\}^p$, which equals 1 if there exist a set $F(i_l)$ among the sets $F(i_1), \ldots, F(i_p)$ that is disjoint from the other sets.

Let

$$C(N,p) \equiv N^{-p/2} \textstyle\sum_{\vec{i}} (1 - R(\vec{i}))$$

Then

$$\parallel X_N(f) \parallel_p^p \leq C(N,p) C^p \parallel f \parallel^p .$$

Proof. We have

$$\begin{array}{l} E\left(1/\sqrt{N} \sum_i f_i\right)^p = N^{-p/2} \sum_{i_1,\ldots,i_p} E \prod_{j=1}^p f_{i_j} \\ = N^{-p/2} \sum_{i_1,\ldots,i_p} (1 - R(i_1, \ldots, i_p)) E \prod_{j=1}^p f_{i_j}. \end{array}$$

By the previous lemma, we have $E \prod_{j=1}^p f_j \leq C^p \parallel f \parallel^p$ for a $C < \infty$, so that we obtain

$$E\left(1/\sqrt{N} \sum_i f_i\right)^p \leq N^{-p/2} \sum_{i_1,\ldots,i_p} (1 - R(i_1, \ldots, i_p)) C^p \parallel f \parallel^p .$$

This completes the proof. □

Lemma 5 *Assume that, for each i, and each p, we have a universal constant C so that*

$$\parallel f_i(f) \parallel_p = (E(f_i(f)(O))^p)^{1/p} \leq C \parallel f \parallel .$$

Assume also that $\max_i \mid F_i \mid \leq K$. *For p an integer, we have* $E \mid X_N(f) \mid^p \leq C(N,p)C^p \parallel f \parallel^p$, *where*

$$\begin{aligned} C(N,p) &\equiv N^{-p/2} \sum_{i_1,\ldots,i_p} (1 - R(i_1,\ldots,i_p)) \\ &\leq 2^p (K^2 p)^{p/2} (N - K^2 p)^{p/2} N^{-p/2}. \end{aligned}$$

For $\psi(x) = e^{x^2} - 1$, *we have* $\parallel X_N(f) \parallel_\psi \leq C_1 \parallel f \parallel$ *for some universal* $C_1 < \infty$.

Proof. We first need to show that $C(N,p) \leq 2^p (K^2 p)^{p/2} (N - K^2 p)^{p/2} N^{-p/2}$. One needs to select p times in a row an element in $\{1,\ldots,N\}$, which then results in one particular $\vec{i}$. Without restrictions on this sequence of p draws, one has N options at each of the subsequent p steps resulting in N^p vectors. Suppose we have arrived at the l-th draw, so that we have a sequence $(i_1,\ldots,i_{l-1})$ with corresponding sets $F(i_1),\ldots,F(i_{l-1})$ For a given set $F(i_s)$ we define the set $F^c(i_s) = \cup_{i \in F(i_s)} F(i)$. For a next i_l we define a binary $B(i_l) = 1$ if $F^c(i_l)$ has no intersection with $\cup_{s \leq l-1} F^c(i_s)$, and we define $B(i_l) = 0$ otherwise. The maximal size of a set $F(i)$ is K so that the maximal size of a set $F^c(i)$ is $K^* = K^2$. Suppose $B(i_l) = 1$. Among $F(i_1),\ldots,F(i_l)$, $F(i_l)$ is an island, and one cannot find a single element in $\{1,\ldots,N\}/\{i_1,\ldots,i_l\}$ that connects both i_l and one of the others in $\{i_1,\ldots,i_{l-1}\}$. In general, an element with $B(i_l) = 1$ will need at least one future $s > l$ selection with $B(i_s) = 0$ in order to connect i_l with i_s. Suppose in the sequence of p elements $(B(i_1),\ldots,B(i_p))$ there are more than $p/2$ 1's. Then there will be at least one island of size 1 one corresponding with a i_l with $B(i_l) = 1$. Thus, in that case $1 - R(\vec{i}) = 0$. Thus, we only need to count the vectors for which $B(i_1),\ldots,B(i_p)$ has at most $p/2$ 1's. For a choice with $B(i_l) = 1$, we have at most $N - K^2 p$ possible choices since we cannot select any of the elements in $F(i_1)^+,\ldots,F(i_{l-1})^+$. For a choice with $B(i_l) = 0$, we have maximally $K^2 p$ choices. The total number of sequences $B(i_1),\ldots,B(i_p)$ for which there are at most $p/2$ 1's is upper-bounded by 2^p. The total number of sequences $\vec{i}$ present in one such sequence is given by $(K^2 p)^{p/2} (N - K^2 p)^{p/2}$. To conclude, we have the following upper bound

$$C(N,p) \leq 2^p (K^2 p)^{p/2} (N - K^2 p)^{p/2} N^{-p/2}.$$

This proves that $E \mid X_N(f) \mid^p \leq C^p C(N,p) \parallel f \parallel^p$. We now want to bound the orlics norm $\parallel X_N(f) \parallel_{\psi_2}$. Firstly, we will do this for the orlics norm

$\psi_1(x) = \exp(x) - 1$. Using that $p! \geq (p/2)!(p/2)^{p/2}$, $(N - K^2p)/N \leq 1$, we have

$$\begin{aligned}
&\| X_N(f) \|_{\psi_1} = \inf\left\{c_0 : \sum_{p=1}^{\infty} \frac{C^p C(p,N)}{p!} \frac{\|f\|^p}{c_0^p} \leq 1\right\}\\
&= \inf\left\{c_0 : \sum_{p=1}^{\infty} \frac{C^p 2^p}{p!} (K^2p)^{p/2} \frac{(N-K^2p)^{p/2}}{N^{p/2}} \frac{\|f\|^p}{c_0^p} \leq 1\right\}\\
&\leq \inf\left\{c_0 : \sum_{p=1}^{\infty} \left(\frac{2C\sqrt{K^2}\|f\|}{c_0}\right)^p \frac{p^{p/2}}{(p/2)!(p/2)^{p/2}} \leq 1\right\}\\
&= \inf\left\{c_0 : \sum_{p=1}^{\infty} \left(\frac{2\sqrt{2}C\sqrt{K^2}\|f\|}{c_0}\right)^p \frac{1}{(p/2)!} \leq 1\right\}.
\end{aligned}$$

Thus there exists a $c_0 = c_0(K^2, C) \parallel f \parallel$ so that the term on the left of the inequality is smaller or equal than 1, so that we have shown $\parallel X_N(f) \parallel_{\psi_1} \leq c_0(K^2, C) \parallel f \parallel$. This completes the proof.

Let's now do the proof for the orlics norm identified by ψ_2. Note $E\exp\{(X_n(f)/c_0)^2\} - 1 = \sum_{p=1}^{\infty} \frac{EX_n(f)^{2p}}{c_0^{2p} p!}$. Thus, we have

$$\begin{aligned}
&\| X_N(f) \|_{\psi_2} = \inf\left\{c_0 : \sum_{p=1}^{\infty} \frac{C^{2p} C(2p,N)}{p!} \frac{\|f\|^{2p}}{c_0^{2p}} \leq 1\right\}\\
&= \inf\left\{c_0 : \sum_{p=1}^{\infty} \frac{C^{2p} 2^{2p}}{p!} (K^2 2p)^p \frac{(N-K^2 2p)^p}{N^p} \frac{\|f\|^{2p}}{c_0^{2p}} \leq 1\right\}\\
&\leq \inf\left\{c_0 : \sum_{p=1}^{\infty} \left(\frac{2\sqrt{2}C\sqrt{K^2}\|f\|}{c_0}\right)^{2p} \frac{p^p}{p!} \leq 1\right\}.
\end{aligned}$$

The term within $()^{2p}$ can be made smaller than an arbitrary number $\delta > 0$ by just selecting c_0 large enough. Therefore, we need to show that $\sum_{p=1}^{\infty} \delta^p p^p / p!$ is bounded for some small enough $\delta > 0$. The proof then proceeds as above for the ψ_1-orlics norm. Now, we note that, using $1 - x \approx \exp(-x)$ for $x \approx 0$, and $\sum_{j=1}^{p} (j-1) = p(p-1)/2$,

$$\begin{aligned}
&\frac{p!}{p^p} = \prod_{j=1}^{p} \{1 - (j-1)/p\} \approx \exp(-\sum_{j=1}^{p} (j-1)/p)\\
&= \exp(-\sum_{j=1}^{p} (j-1)/p) = \exp(-1/pp(p-1)/2) = \exp(-(p-1)/2).
\end{aligned}$$

Thus, $\sum_{p=1}^{\infty} \delta^p p^p / p!$ behaves as $\sum_{p=1}^{\infty} \delta^p \exp((p-1)/2)$. Since $\exp((p-1)/2) \leq \exp(p)$ is bounded by a p-th power of e, by selecting δ small enough, this sum can be made arbitrarily small. This completes the proof.

□

B.4 (A3): Asymptotic equicontinuity of Z_Y^N.

The process $Z_Y^N = 1/\sqrt{N} \sum_i f_i^Y$ is a sum of independent random variables conditional on (W, A), so that its analysis is a simple imitation of the general analysis above. The proof that the $\parallel \cdot \parallel_p$ norm of $f_i^Y(f)$ is bounded by the supremum norm of f is trivial.

B.5 (A3): Asymptotic equicontinuity of Z_A^N

Conditional on W, for a fixed $f = (\tilde{h}, \bar{Q})$, we can represent this process as $1/\sqrt{N}\sum_i f_i^A(A_j : j \in F_i) = 1/\sqrt{N}\sum_{i=1}^N f_i^A(C_i(A_j : j \in F_i))$, $E(f_i^A \mid W) = 0$, where $C_i \in \mathbb{R}^d$ for some d so that we can define $f_i^A : \mathbb{R}^d \rightarrow \mathbb{R}$ for a fixed d not depending on N, and A_i, $i = 1, \ldots, N$ are independent. Again, the above general analysis can be applied, and the proof that the $\| \cdot \|_p$-norm of $f_i^A(f)$ is bounded by supremum norm of f is trivial.

B.6 (A3): Asymptotic equicontinuity of $Z_W^N(\tilde{h}, \bar{Q})$.

Conditional on $F_1, \ldots, F_N$, we can represent $Z_W^N(\tilde{h}, \bar{Q})$ as $1/N\sum_i\{f_i^W(\tilde{h}, \bar{Q})(W_j : j \in R_i) - P_0 f_i\}$, where W_j, $j = 1, \ldots, N$, are independent. Specifically, $f_i^W(\tilde{h}, \bar{Q}) = \int \tilde{h}(\bar{Q} - \bar{Q}_0)g_{0,i}(c \mid W)$. If $F_i \cap F_j = \emptyset$, then, conditional on $F_1, \ldots, F_N$, f_i and f_j are not only uncorrelated, but are completely independent. Thus, we can apply our general proof above to establish the bound of its orlics norm. It is again trivially shown that the $\| \cdot \|_p$ norm of $f_i^W(f)$ is bounded by the supremum norm of f.

C Proof of (A2).

Define the process $Z_{W,N}^1(\bar{Q}) = \frac{1}{\sqrt{N}}\sum_{i=1}^N\{f_{W,i}^1(\bar{Q}) - P_0 f_{W,i}^1(\bar{Q})\}$ indexed by $\bar{Q}$, where $f_{W,i}^1(\bar{Q}) = \int \bar{Q}(c)g_i^*(c \mid W)$. We need to prove that $R_{N,0} = Z_{W,N}^1(\bar{Q}_N^* - \bar{Q}^*) = o_P(1)$. This proof is completely analogue to our proof above for establishing asymptotic equicontinuity of the other $Z_{W,N}(\tilde{h}, \bar{Q})$ process analyzed above, but now with respect to athe supremum norm for $\bar{Q}$.

D (A1): Establishing weak convergence of first order approximation of standardized estimator

D.1 Outline of proof.

Recall

$$\psi_N^* - \psi_0 \approx \sum_i f_i(O) = \sum_i\{f_{Y,i} + f_{A,i} + f_{W,i}\} = Z_Y^N + Z_A^N + Z_W^N,$$

where

$$\begin{aligned}
f_{Y,i} &= \frac{\bar{h}(g^*)}{\bar{h}(g_0)}(c_i^Y)(Y_i - \bar{Q}_{Y,0}(c_i^Y)) \\
f_{A,i} &= \frac{\bar{h}(g^*)}{\bar{h}(g_0)}(c_i^Y)(\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c_i^Y) \\
&\quad - \int_c \frac{\bar{h}(g^*)}{\bar{h}(g_0)}(c)(\bar{Q}_{Y,0} - \bar{Q}_Y^*)(c) g_{0,i}(c \mid W) \\
f_{W,i} &= \int_c \bar{Q}_{Y,0}(c) g_i^*(c \mid W) - \int_{c,w} \bar{Q}_{Y,0}(c) g_i^*(c \mid W) Q_{W,0}(w).
\end{aligned}$$

We will establish weak convergence of each of the three terms separately.

The proof of weak convergence of Z_Y^N can be based on standard CLT since, conditional on (A, W), Z_Y^N is a sum of mean zero independent random variables.

Lemma 6 $Z_Y^N = 1/\sqrt{N}\sum_{i=1}^N f_{Y,i}$ *converges weakly to a normal distribution with mean zero and variance*

$$\begin{aligned}
\sigma_Y^2 &= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N P_0 f_{Y,i}^2 \\
&= \lim_{N\to\infty} \int \tilde{h}_0(c)^2 \sigma_Y^2(c) \frac{\bar{h}_0(c)}{N} d\mu(c),
\end{aligned}$$

where

$$\sigma_Y^2(c_i^Y) = E_0(\{Y_i - \bar{Q}_{Y,0}(c_i^Y)\}^2 \mid A, W) = E_0(\{Y_i - \bar{Q}_{Y,0}(c_i^Y)\}^2 \mid c_i^Y(A, W)).$$

For example, if Y_i is binary, then the latter expression equals

$$\sigma_Y^2(c_i^Y) = \bar{Q}_{Y,0}(1 - \bar{Q}_{Y,0})(c_i^Y).$$

Recall that $\bar{h}_0/N = \frac{1}{N}\sum_i h_{0,i}$.

We establish weak convergence of Z_A^N by establishing convergence of its p-th moment. Specifically, we establish that $E(Z_A^N)^p \approx \bar{\rho}^{p/2}\frac{p!}{(p/2)!2^{p/2}}$ for p even, and $E(Z_A^N)^p \approx 0$ for p odd, where $\bar{\rho}$ represents the limit of the second moment $E(Z_A^N)^2$. This convergence in moments implies that Z_N converges weakly to a normal distribution $N(0, \sigma^2 = \bar{\rho})$, where we utilize the following two lemmas.

Lemma 7 *A random variable Z with $EZ^p = \bar{\rho}^{p/2}\frac{p!}{2^{p/2}(p/2)!}$ for p even, and $EZ^p = 0$ for p odd has probability distribution equal to $N(0, \sigma^2 = \bar{\rho})$, the normal distribution with mean zero and variance $\bar{\rho}$.*

Proof. We have

$$\begin{aligned}
&E\exp(tZ) = \textstyle\sum_{p=0}^{\infty} \frac{t^p}{p!} EZ^p = \sum_{p=0}^{\infty} \frac{t^{2p}}{(2p)!} EZ^{2p} \\
&= \textstyle\sum_{p=0}^{\infty} \frac{t^{2p}}{(2p)!} \bar{\rho}^p \frac{(2p)!}{p!2^p} \\
&= \textstyle\sum_{p=0}^{\infty} \frac{(0.5t^2\bar{\rho})^p}{p!} \\
&= \exp(0.5t^2\bar{\rho}),
\end{aligned}$$

which is the moment generating function of $N(0, \sigma^2 = \bar{\rho})$, i.e., a normal distribution with mean zero and variance equal to $\bar{\rho}$. □

Lemma 8 *Suppose* $EZ_N^p \leq C^p$ *for a universal* $C < \infty$. *Suppose that* $EZ_N^p \rightarrow \bar{\rho}^{p/2} p!/(p/2)!2^{p/2}$ *for* p *even, and* $EZ_N^p \rightarrow 0$ *for* p *odd. Then* Z_N *converges in distribution to* $Z = N(0, \sigma^2 = 2d)$.

Proof. Consider the moment generating function $E\exp(tZ_N)$ when $EZ_N^p \rightarrow \bar{\rho}^{p/2} \frac{p!}{(p/2)!2^{p/2}}$. By Fubini's theorem,

$$E \textstyle\sum_{p=0}^{\infty} \frac{t^p}{(p)!} Z_N^p = \sum_{p=0}^{\infty} \frac{t^p}{p!} EZ_N^p.$$

Because $EZ_N^p \leq C^p$, we have

$$\sum_{p=M}^{\infty} \frac{t^p}{p!} EZ_N^p \leq \sum_{p=M}^{\infty} \frac{t^p}{p!} C^p,$$

which converges to zero in $M \rightarrow \infty$. Therefore, we can truncate the summation defining the moment generating function of Z_N and focus on establishing convergence of $E \sum_{p=0}^{M} \frac{t^p}{p!} Z_N^p$, but the latter follows from $EZ_N^p \rightarrow EZ^p$ as $N \rightarrow \infty$. This proves that

$$E\exp(tZ_N) \rightarrow E\exp(tZ).$$

This proves that $Z_N(Q)$ converges in distribution to $Z(Q) = N(0, \sigma^2 = \bar{\rho})$ as $N \rightarrow \infty$. □

D.2 (A1): Establishing convergence of p-th moment for Z_A^N

Lemma 9 *Let* $Z_A^N = \sum_i f_i(A)$, *and* $f_i(A) = f_i(A_j : j \in F_i)$. *Let*

$$\rho(j_1, j_2 \mid W) = E_0(f_{j_1}(A) f_{j_2}(A) \mid W).$$

Specifically, for Z_A^N we have

$$\begin{aligned}&\rho(j_1,j_2\mid W)=\textstyle\int \frac{\bar h_0^*(c_1)}{\bar h_0(c_1)}(\bar Q_0-\bar Q)(c_1)\frac{\bar h_0^*}{\bar h_0}(c_2)(\bar Q_0-\bar Q)(c_2)g_{0,j_1,j_2}(c_1,c_2\mid W)\\ &-\textstyle\int \frac{\bar h_0^*}{\bar h_0}(c)(\bar Q_0-\bar Q)(c)g_{0,j_1}(c\mid W)\int \frac{\bar h_0^*}{\bar h_0}(c)(\bar Q_0-\bar Q)(c)g_{0,j_2}(c\mid W),\end{aligned}$$

where $g_{0,i,j}$ is conditional distribution of $(C_i(A,W),C_j(A,W))$, given W, which only depends on A through $(A_l : l\in F_i\cup F_j)$. Let $\rho_A(j_1,j_2)=E(\rho(j_1,j_2\mid W)\mid \mathbf{F})$. For two integers (i_1,i_2), define $R_2(i_1,i_2)$ as the indicator that the intersection of $N(i_1)$ and $N(i_2)$ is non-empty. Assume

$$\frac{1}{N}\sum_{i_1,i_2,R_2(i_1,i_2)=1}\rho_A(i_1,i_2)\rightarrow_{N\rightarrow\infty}\bar\rho_A.$$

We have for p even,

$$E\left(\tfrac{1}{\sqrt{N}}\textstyle\sum_i f_i\right)^p\rightarrow \tfrac{p!}{(p/2)!2^{p/2}}\bar\rho_A^{p/2} \text{ as } N\rightarrow\infty.$$

For p odd, this p-th moment converges to zero.

Proof. Given an index $\vec{i}=(i_1,\ldots,i_p)\in\{1,\ldots,N\}^p$ (one among N^p), we can draw a graph by drawing a line between two elements i_{l_1},i_{l_2} in $\{i_1,\ldots,i_p\}$ whenever the two corresponding sets $F(i_{l_1})$ and $F(i_{l_2})$ have a non-empty intersection. Classify an element $(i_1,\ldots,i_p)$ by the sizes of the connected sets that make up the graph of $(i_1,\ldots,i_p)$. One category of indices is that each connected set is of size 2, assuming p is even, and let $R_2(\vec{i})$ be the indicator of falling in this category. For each of the other categories with all connected sets of size *larger or equal than 2, but at least one larger than 2*, we assume that its number X of elements is of smaller order than $N^{-p/2}$: $N^{-p/2}X\rightarrow 0$ as $N\rightarrow\infty$. This can be shown to hold under our assumption that $\mid F_i\mid< K$. In addition, for $\vec{i}$ with $R_2(\vec{i})=1$, let $j=1,\ldots,p/2$ index the $p/2$ pairs that are connected, and let $j_1(\vec{i}),j_2(\vec{i})$ denote the two indices in $\{i_1,\ldots,i_p\}$ corresponding with each j-th pair. We also note that (f_{j_1},f_{j_2}) are independent across the pairs j, conditional on W. We can now state

$$\begin{aligned}&E\left(\tfrac{1}{\sqrt{N}}\textstyle\sum_i f_i\right)^p=N^{-p/2}\textstyle\sum_{i_1,\ldots,i_p}R_2(i_1,\ldots,i_p)\prod_{j=1}^{p/2}Ef_{j_1(\vec{i})}f_{j_2(\vec{i})}+o(1)\\ &=N^{-p/2}\textstyle\sum_{i_1,\ldots,i_p}R_2(i_1,\ldots,i_p)\prod_{j=1}^{p/2}\rho(j_1,j_2\mid W)+o(1),\end{aligned}$$

where

$$\rho(j_1,j_2\mid W)=\int\frac{h(g^*)(c_1)}{h(g_0)(c_1)}(\bar Q_0-\bar Q)(c_1)\frac{h(g^*)}{h(g_0)}(c_2)(\bar Q_0-\bar Q)(c_2)g_{0,j_1,j_2}(c_1,c_2\mid W).$$

Let (F_i, W_i) represent the i-specific baseline covariates, so that F_i is separate from W_i. We take a conditional expectation, given $F_1, \ldots, F_N$. Conditional on $F_1, \ldots, F_N$, the indicators $R_2(\vec{i})$ are fixed. Since $\rho(j_1, j_2 \mid W)$ only depends on W through $(W_i : i \in N_{j_1} \cup N_{j_2})$, the sets $N_{j_1} \cup N_{j_2}$ in the product over j are disjoint across j, and for any pair (t, l), W_t, W_l is conditionally independent, it follows that, conditional on $\mathbf{F} = (F_1, \ldots, F_N)$,

$$E\left(\frac{1}{\sqrt{N}} \sum_i f_i(\bar{Q})\right)^p \approx N^{-p/2} \sum_{i_1,\ldots,i_p} R_2(i_1, \ldots, i_p) \prod_{j=1}^{p/2} E(\rho(j_1, j_2 \mid W) \mid \mathbf{F}).$$

Let $\rho_N(j_1, j_2) = E(\rho(j_1, j_2 \mid W) \mid \mathbf{F})$. For two integers (i_1, i_2), define $R_2(i_1, i_2)$ as the indicator that the intersection of $N(i_1)$ and $N(i_2)$ is non-empty. Let $\mathcal{R}_2 = \{(i_1, i_2) \in \{1, \ldots, N\}^2 : R_2(i_1, i_2) = 1\}$, and $\mathcal{R}_2^{p/2}$ is the cartesian product of this set. Let $\mathcal{R} = \{(i_1, \ldots, i_p) : R_2(\vec{i}) = 1\}$, where we are reminded that $R_2(\vec{i})$ is the indicator of all connected sets among $\{i_1, \ldots, i_p\}$ being of size 2. We have the following lemmas.

Lemma 10 *We have*

$$N^{-p/2} \sum_{(j_1,j_2) \in \mathcal{R}_2^{p/2}} R_2((j_1, j_2 : j = 1, \ldots, p/2) \prod_{j=1}^{p/2} \rho_N(j_1, j_2)$$
$$= N^{-p/2} \sum_{(i_1,i_2) \in \mathcal{R}_2^{p/2}} \prod_{j=1}^{p/2} \rho_N(j_1, j_2) + o(1).$$

Proof of Lemma 10. Note that the right-hand side sums over vectors $\mathcal{R}_2^{p/2}$ while the left-hand side sums over vectors that are both in $\mathcal{R}_2^{p/2}$ *and* satisfy $\vec{i} \in \mathcal{R}$. Since a vector made up of $p/2$-connected pairs can correspond with connected sets of larger size than 2, we have that $\mathcal{R} \subset \mathcal{R}_2^{p/2}$, i.e., the right-hand side sums over more elements. However, the number of these extra vectors $\vec{i} \in \mathcal{R}_2^{p/2}/\mathcal{R}$ that should not have been counted is of smaller order than $N^{p/2}$, so that the contribution is negligible. □

Lemma 11 *We have*

$$N^{-p/2} \sum_{i_1,\ldots,i_p} R_2(i_1, \ldots, i_p) \prod_{j=1}^{p/2} \rho_N(j_1, j_2)$$
$$= \frac{p!}{(p/2)! 2^{p/2}} N^{-p/2} \sum_{(j_1,j_2) \in \mathcal{R}_2^{p/2}} R_2(j_1, j_2 : j = 1, \ldots, p/2) \prod_{j=1}^{p/2} \rho_N(j_1, j_2).$$

Proof of Lemma 11: Consider a vector of three connected pairs $(1, 1), (2, 2), (3, 3)$ (i.e., $p = 6$). These three connected pairs appear 3! (i.e. $(p/2)!$) times on right-hand side. However, on the left-hand side, any vector of length 6 with two 1's, two 2's, and 2 3's is counted, and there are $6!/2^3$ (i.e., $p!/2^{p/2}$) of such vectors: the number of ordered vectors of length 6 is 6!, but flipping the two 1's or two 2's or two 3's does not yield a different vector. □
Finally, we state the following trivial result

Lemma 12 *We have*

$$\sum_{(j_1,j_2)\in\mathcal{R}_2^{p/2}} \prod_{j=1}^{p/2} \rho_N(j_1,j_2) = \left(\sum_{i_1,i_2,R_2(i_1,i_2)=1} \rho_N(i_1,i_2) \right)^{p/2} .$$

This proves that

$$\begin{array}{l} N^{-p/2}\sum_{i_1,\ldots,i_p} R_2(i_1,\ldots,i_p)\prod_{j=1}^{p/2}\rho_N(j_1,j_2) = \\ \frac{p!}{(p/2)!2^{p/2}} N^{-p/2}\sum_{(j_1,j_2)\in\mathcal{R}_2^{p/2}} R_2(j_1,j_2 : j=1,\ldots,p/2)\prod_{j=1}^{p/2}\rho_N(j_1,j_2) \\ \approx \frac{p!}{(p/2)!2^{p/2}} N^{-p/2}\sum_{(j_1,j_2)\in\mathcal{R}_2^{p/2}} \prod_{j=1}^{p/2}\rho_N(j_1,j_2) \\ = \frac{p!}{(p/2)!2^{p/2}} \left(1/N\sum_{i_1,i_2,R_2(i_1,i_2)=1}\rho_N(i_1,i_2)\right)^{p/2} . \end{array}$$

Finally, we assumed that the latter summation within the power converges to $\bar{\rho}$. Thus,

$$E\left(\frac{1}{\sqrt{N}}\sum_i f_i(\bar{Q})\right)^p \rightarrow \frac{p!}{(p/2)!2^{p/2}}\bar{\rho}^{p/2} \; \Box$$

D.3 (A1): Convergence of p-th moment of Z_W^N.

The same proof can be applied to establish the convergence of the p-th moment of Z_W^N resulting in the following lemma.

Lemma 13 *Let* $Z_W^N = \sum_i (f_i(W) - P_0 f_i)$*, and* $f_i(W) = f_i(W_j : j \in R_i)$ *for set* R_i *defined by* $\mathbf{F}$ *with* $\mid R_i \mid < K$ *for some fixed* $K < \infty$*, where we condition on* $\mathbf{F}$*. Let*

$$\rho_W(j_1,j_2) = E_0(f_{j_1}(W)f_{j_2}(W) \mid \mathbf{F}) - E_0(f_{j_1}(W)\mid\mathbf{F})E_0(f_{j_2}(W)\mid\mathbf{F}).$$

Specifically, for Z_W^N *we have*

$$f_i(W) = \int \bar{Q}_{Y,0}(c_i^Y(a,W))g^*(a \mid W) = \int \bar{Q}_{Y,0}(c)g_{0,i}^*(c \mid W).$$

We assumed that $g^*((A_j : j \in F_i) \mid W)$ *only depends on* $(W_j : j \in R_i)$ *for sets* R_i *implied by* $\mathbf{F}$*. Thus, in this case*

$$\begin{array}{l} \rho_W(j_1,j_2) = E_W \int \bar{Q}_{Y,0}(c_1)\bar{Q}_{Y,0}(c_2)g_{0,j_1}(c \mid W)g_{0,j_2}(c \mid W) \\ -E_W \int \bar{Q}_{Y,0}(c)g_{0,j_1}(c\mid W)E_W\int\bar{Q}_{Y,0}(c)g_{0,j_2}(c \mid W). \end{array}$$

For two integers (i_1, i_2), define $R_2(i_1, i_2)$ as the indicator that the intersection of F_{i_1} and F_{i_2} is non-empty. Assume

$$\frac{1}{N} \sum_{i_1, i_2, R_2(i_1, i_2)=1} \rho_W(i_1, i_2) \rightarrow_{N \rightarrow \infty} \bar{\rho}.$$

We have for p even,

$$E\left(\tfrac{1}{\sqrt{N}} \textstyle\sum_i f_i\right)^p \rightarrow_{N \rightarrow \infty} \tfrac{p!}{(p/2)! 2^{p/2}} \bar{\rho}^{p/2}.$$

For p odd, this p-th moment converges to zero.

E Outline of proof of asymptotic normality of TMLE for general longitudinal data.

Recall that $D^* = 1/N \sum_{j=1}^N D_j^*(O)$ is a sum over the individuals j. We will use the notation $P_N D^* = 1/N \sum_{j=1}^N D_j^*(O)$, while $P_0 D^* = 1/N \sum_j E_{P_0} D_j^*(O)$ is its expectation w.r.t. distribution P_0. Let $W = L(0)$. We have $D^* = D_W^* + D_1^*$, $P_0 D_1^*(Q_N^*, Q_{W,0}) = 0$, $P_0 D_1^*(Q_N^*, Q_{W,0}) = \psi_0 - \Psi(Q_N^*, Q_{W,0}) + R_N$, and $P_N D_W^*(Q_N^*, Q_{W,N}) = P_N D_1^*(Q_N^*, Q_{W,N}) = 0$. We assume that $R_N = o_P(1/\sqrt{N})$. In particular, this yields

$$P_0 D^*(Q_N^*, Q_{W,0}) = \psi_0 - \Psi(Q_N^*, Q_{W,0}) + R_N.$$

We now proceed as follows:

$$\begin{aligned}
&\Psi(Q_N^*, Q_{W,N}) - \psi_0 = \Psi(Q_N^*, Q_{W,N}) - \Psi(Q_N^*, Q_{W,0}) + \Psi(Q_N^*, Q_{W,0}) - \psi_0 \\
&= \Psi(Q_N^*, Q_{W,N}) - \Psi(Q_N^*, Q_{W,0}) - P_0 D_1^*(Q_N^*, Q_{W,0}) + R_N \\
&= \Psi(Q_N^*, Q_{W,N}) - \Psi(Q_N^*, Q_{W,0}) + (P_N - P_0) D_1^*(Q_N^*, Q_{W,0}) \\
&\qquad\qquad - P_N\{D_1^*(Q_N^*, Q_{W,0}) - D_1^*(Q_N^*, Q_{W,N})\} + R_N \\
&= \Psi(Q_N^*, Q_{W,N}) - \Psi(Q_N^*, Q_{W,0}) + (P_N - P_0) D_1^*(Q_N^*, Q_{W,0}) \\
&\qquad\qquad + (P_N - P_0)\{D_1^*(Q_N^*, Q_{W,N}) - D_1^*(Q_N^*, Q_{W,0})\} \\
&\qquad\qquad + P_0\{D_1^*(Q_N^*, Q_{W,N}) - D_1^*(Q_N^*, Q_{W,0})\} + R_N.
\end{aligned}$$

Assume,

$$\begin{aligned}
&\Psi(Q_N^*, Q_{W,N}) - \Psi(Q_N^*, Q_{W,0}) \approx \Psi(Q^*, Q_{W,N}) - \Psi(Q^*, Q_{W,0}) \\
&= 1/N \textstyle\sum_i f_{W,i}^1 + o_P(1/\sqrt{N}) \\
&(P_N - P_0)\{D_1^*(Q_N^*, Q_{W,N}) - D_1^*(Q_N^*, Q_{W,0})\} = o_P(1/\sqrt{N}) \\
&P_0\{D_1^*(Q_N^*, Q_{W,N}) - D_1^*(Q_N^*, Q_{W,0})\} = P_0\{D_1^*(Q^*, Q_{W,N}) - D_1^*(Q^*, Q_{W,0})\} \\
&\qquad\qquad + o_P(1/\sqrt{N}) \\
&= 1/N \textstyle\sum_i f_{W,i}^2 + o_P(1/\sqrt{N}) \\
&(P_N - P_0) D_1^*(Q_N^*, Q_{W,0}) = (P_N - P_0) D_1^*(Q^*, Q_{W,0}) + o_P(1/\sqrt{N}).
\end{aligned}$$

To establish these $o_P(1/\sqrt{N})$-results, we will have to establish asymptotic equicontinuity of a process $Z_N(h,Q)$ indexed by (h,Q), and we can use the same template as we used for the single time point case to analyze these processes. Thus, any function $f(O)$ is decomposed in $\sum_{k=0}^{\tau+1} f_{L(k)} + \sum_{k=0}^{\tau} f_{A(k)}$, where $f_{L(k)} = E_0(f(O) \mid \bar{L}(k), \bar{A}(k-1)) - E_0(f(O) \mid \bar{L}(k-1), \bar{A}(k-1))$, and $f_{A(k)} = E_0(f(O) \mid \bar{L}(k), \bar{A}(k)) - E_0(f(O) \mid \bar{L}(k), \bar{A}(k-1))$. In this manner, we obtain an orthogonal decomposition of $Z_N(h,Q) = \sum_{k=0}^{\tau+1} Z^N_{L(k)}(h,Q) + \sum_{k=0}^{\tau} Z^N_{A(k)}(h,Q)$. Each of these orthogonal components of this process will be analyzed separately completely analogue as we analyzed our orthogonal components $Z_{N,Y}$, $Z_{N,A}$ and $Z_{N,W}$ for the single time point case. For example, we can represent $Z_{N,L(k)}$ as $\sum_i f_i(L(k) \mid Pa(L(k)))$, where f_i has conditional mean zero, given $Pa(L(k)) = (\bar{L}(k-1), \bar{A}(k-1))$, and f_i only depends on $L(k)$ through $L_j(k)$ for j in a finite set (e.g., $F_i(k)$), and apply our general results above for analyzing such processes.

In this manner, we obtain the following first order approximation

$$\begin{aligned}\psi^*_N - \psi_0 \;\; = \;\; & \frac{1}{N}\sum_{i=1}^{N}\{f^1_{W,i}(W) - P_0 f^1_{W,i}\} + \frac{1}{N}\sum_i \{f^2_{W,i}(W) - P_0 f^2_{W,i}\} \\ & + (P_N - P_0) D^*_1(Q^*, Q_{W,0}) + o_P(1/\sqrt{N}).\end{aligned}$$

Analysis of first order approximation: Let $\bar{f}_{W,i} = f^1_{W,i} + f^2_{W,i}$. The first order approximation equals

$$\begin{aligned}& 1/N \textstyle\sum_i \{D^*_{1,i}(Q^*, Q_{W,0}) + \bar{f}_{W,i}(W) - P_0\{D^*_{1,i} + \bar{f}_{W,i}\}\} \\ & \equiv 1/N \textstyle\sum_i f_i(O).\end{aligned}$$

Using our notation, this thus equals $(P_N - P_0)f$. It remains to prove that this first order expansion converges to a normal limit distribution. This proof has its own outline. Firstly, we decompose any $f(O)$ as $f = 1/N \sum_{k=0}^{\tau+1} f_{L(k)}(O) + \sum_{k=0}^{\tau} f_{A(k)}(O)$, where

$$\begin{aligned}f_{L(k)}(O) &= E(f \mid \bar{L}(k), \bar{A}(k-1)) - E(f \mid \bar{L}(k-1), \bar{A}(k-1)) \\ f_{A(k)}(O) &= E(f \mid \bar{L}(k), \bar{A}(k)) - E(f \mid \bar{L}(k), \bar{A}(k-1)).\end{aligned}$$

In this manner, we decompose $\sqrt{N}(P_N - P_0)f$ orthogonally in $\sum_{k=0}^{\tau+1} Z^N L(k)(f) + \sum_{k=0}^{\tau} Z^N_{A(k)}(f)$, where $Z^N_{L(k)} = 1/\sqrt{N}\sum_{i=1}^N f_{i,L(k)}$ and $Z^N_{A(k)}(f) = 1/\sqrt{N}\sum_{i=1}^N f_{i,A(k)}$.

We will then need to show that

$$\begin{aligned}Z^N_{L(k)} &\Rightarrow_d N(0, \sigma^2_{L(k)}) \\ Z^N_{A(k)} &\Rightarrow_d N(0, \sigma^2_{A(k)}).\end{aligned}$$

Due to the orthogonality of the empirical processes, using moment generating functions, it also follows that

$$\sum_k Z^N_{L(k)} + \sum_k Z^N_{A(k)} \Rightarrow_d N(0, \sigma^2 = \sum_k \sigma^2_{L(k)} + \sum_k \sigma^2_{A(k)}).$$