Identification and Estimation of Causal Effects Defined by Shift Interventions

Numair Sani
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218
snumair1@jhu.edu

Jaron J. R. Lee
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218
jaron.lee@jhu.edu

Ilya Shpitser
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218
ilyas@cs.jhu.edu

Abstract

Causal inference quantifies cause effect relationships by means of counterfactual responses had some variable been artificially set to a constant. A more refined notion of manipulation, where a variable is artificially set to a fixed function of its natural value is also of interest in particular domains. Examples include increases in financial aid, changes in drug dosing, and modifying length of stay in a hospital.

We define counterfactual responses to manipulations of this type, which we call shift interventions. We show that in the presence of multiple variables being manipulated, two types of shift interventions are possible. Shift interventions on the treated (SITs) are defined with respect to natural values, and are connected to effects of treatment on the treated. Shift interventions as policies (SIPs) are defined recursively with respect to values of responses to prior shift interventions, and are connected to dynamic treatment regimes. We give sound and complete identification algorithms for both types of shift interventions, and derive efficient semi-parametric estimators for the mean response to a shift intervention in a special case motivated by a healthcare problem. Finally, we demonstrate the utility of our method by using an electronic health record dataset to estimate the effect of extending the length of stay in the intensive care unit (ICU) in a hospital by an extra day on patient ICU readmission probability.

1 INTRODUCTION

Establishing cause effect relationships is a fundamental goal in data-driven empirical science and decision making. An influential approach to causal inference quantifies causal effects by means of responses to an intervention operation, which manipulates variables to attain specified values, possibly contrary to fact. This intervention operation is denoted by do(.) in (Pearl, 2009), and is used to define potential outcome random variables in wide use in statistics and public health (Neyman, 1923; Rubin, 1976).

Other kinds of intervention operations have been considered in the literature. *Dynamic treatment regimes* (DTRs), used in precision medicine and related applications (Chakraborty and Moodie, 2013), manipulate variables to values that depend on causally prior variables. *Edge and path interventions* (Shpitser and Tchetgen Tchetgen, 2016) manipulate variables to distinct values with respect to different causal pathways the variables are involved in. These interventions have been used to quantify direct, indirect, and path-specific effects in mediation analysis. *Soft interventions* (Eberhardt, 2014) "nudge" variables (or the data-generating process for variables) away from their natural state, rather than manipulating them to attain specific constant values. A recent type of intervention of this sort that manipulates the propensity score was considered in (Kennedy, 2019).

In this paper we consider a particular type of soft intervention where variables are manipulated to attain values given by fixed functions of their existing values. We call such interventions *shift interventions*. Shift interventions arise in settings where the counterfactual change of interest is most naturally expressed in terms of existing realizations of variables to be manipulated. Examples of such settings include changes in drug dosing, increases in financial aid, or policy deviations from an existing standard in medical, social, or economic domains. We show that in the presence of multiple variables being manipulated, two types of shift interventions are possible. *Shift interventions on the treated* (SITs) are defined with respect to their naturally observed values, and are connected to *effects of treatment on the treated* (ETTs) (Sh-

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

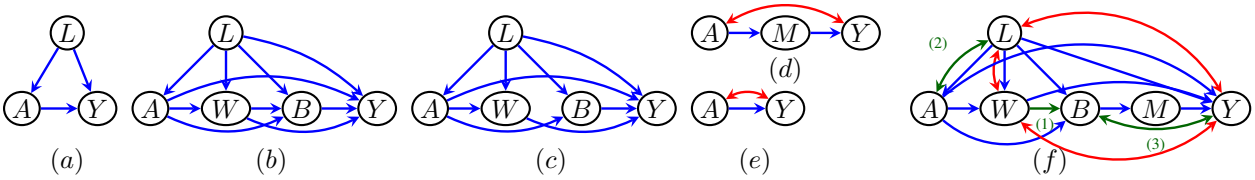


Figure 1: (a) A causal graph representing a single treatment cross-sectional study. (b) A causal graph representing a two stage observational study, with a first line treatment A, and a second line treatment B. (c) A causal graph representing a two stage observational study, where the second line treatment B is not assigned using intermediate outcomes W. (d) The latent projection representing the front-door causal model. (e) The latent projection representing the bow arc causal model. (f) A class of latent projections representing failure of identification of SITs in Theorem 4.

pitser and Pearl, 2009). *Shift interventions as policies* (SIPs) are defined recursively with respect to values of responses to prior shift interventions, and are connected to dynamic treatment regimes. Despite these connections, responses to shift interventions are distinct types of counterfactuals, and we show their identification gives rise to subtleties not present in identification of either DTRs or ETTs.

We give sound and complete identification algorithms for both types of shift interventions, and derive an efficient semi-parametric estimator for the response to a shift intervention in a special case motivated by a healthcare problem. Finally, we demonstrate the utility of our method by using an electronic health record dataset to estimate the effect of extending the length of stay in the intensive care unit (ICU) in a hospital by an extra day on patient ICU readmission probability.

2 PRELIMINARIES

Causal inference aims to establish a link between observed random variables $V \equiv \{V_1, \ldots, V_k\}$ and counterfactual random variables $V_i(a)$, which denote the response of V_i had variables $A \subseteq V$ been manipulated, possibly contrary to fact, to obtain values a. The distribution over $V_i(a)$ is denoted by $p(V_i|\text{do}(a))$ in (Pearl, 2009). Counterfactuals quantify causal effects as contrasts defined by two manipulations, representing treatment and control arms of a hypothetical randomized controlled trial (RCT). For example, the average causal effect (ACE) is defined as $\mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$, where Y is an outcome variable, and A are one or more treatment variables manipulated to values a representing treatments of interest, or values a' representing the baseline treatments in the control group.

An elegant formalism for defining causal models uses directed acyclic graphs (DAGs). A DAG is a directed graph with no directed cycles. In a causal model represented by a DAG $\mathcal{G}$, each vertex in $\mathcal{G}$ corresponds to a variable (we will use the same letter, e.g. V_i, for both). For each V_i, the set of variables with directed arrows into V_i in $\mathcal{G}$, denoted as parents of V_i, or $\text{pa}(V_i)$, is the set of direct causes of V_i, in the following sense. We assume the existence of atomic counterfactual random variables of the form $V_i(a)$, for each value set a in $\mathfrak{X}_{\text{pa}(V_i)}$, the state space of $\text{pa}(V_i)$. We use these random variables to define other counterfactuals by means of the *recursive substitution* definition. For any $V_i \in V, A \subseteq V$, we have

$$V_i(a) \equiv V_i(a_{\text{pa}(V_i)}, \{W(a) : W \in \text{pa}(V_i) \setminus A\}). \quad (1)$$

As an example, consider the DAG $\mathcal{G}$ in Fig. 1 (a), representing an observational study with a single treatment A (representing a drug dose), an outcome Y, and a vector of baseline covariates L. Given $\mathcal{G}$, atomic counterfactuals are of the form L, $A(l)$, $Y(a, l)$, for any values a, l in $\mathfrak{X}_{\{A,L\}}$. Further, we define $Y(a)$ using (1) as $Y(a, L)$. Note that (1) allows definitions of the form $A(a, L) \equiv A(L) \equiv A$, since $A \notin \text{pa}(A)$, and $A \notin \text{pa}(L)$.

Causal models are defined by restrictions on counterfactual random variables. We will work with a popular model called the *structural causal model* (Pearl, 2009), which asserts the following marginal independences:

$$\{V_1(a_1) : a_1 \in \mathfrak{X}_{\text{pa}(V_1)}\} \perp\!\!\!\perp \ldots \perp\!\!\!\perp \{V_k(a_k) : a_k \in \mathfrak{X}_{\text{pa}(V_k)}\}.$$

In our example, these assert

$$L \perp\!\!\!\perp \{A(l) : l \in \mathfrak{X}_L\} \perp\!\!\!\perp \{Y(a, l') : a, l' \in \mathfrak{X}_{\{A,L\}}\}.$$

Causal models such as the structural causal model allow counterfactual quantities such as $p(Y(a))$ to be expressed in terms of the observed data distribution. If all variables in the causal model are observed, every $p(V(a))$ is identified by the following functional:

$$p(V(a)) = \prod_{V_i \in V} p(V_i | \text{pa}(V_i) \setminus A, a_{\text{pa}(V_i)}), \quad (2)$$

known as the extended g-formula. If A is empty, we have

$$p(V) = \prod_{V_i \in V} p(V_i | \text{pa}(V_i)), \quad (3)$$

which is the well-known Bayesian network factorization of the observed data distribution $p(V)$ (Pearl, 1988). In other words, assuming a causal model on a DAG $\mathcal{G}$ implies the observed data distribution $p(V)$ factorizes according to $\mathcal{G}$ as in (3), and all interventional distributions $p(V(a))$ are identified by modified versions, as in (2), of this factorization.

In our example, $p(V(a)) = p(Y(a), A, L)$ is identified as $p(Y|a, L)p(A|L)p(L)$. If we are only interested in $p(Y(a))$, we simply marginalize the modified factorization appropriately to yield the *adjustment formula* $\sum_L p(Y|a, L)p(L)$.

Generalized Interventions and Targets of Inference

Before describing shift interventions, we consider two related counterfactual quantities considered in the causal inference literature. Aside from the example in Fig. 1 (a), we will also consider the causal model in Fig. 1 (b) representing an observational study with two treatments A, B given in stages. A is given based on a set of baseline characteristics L, and would represent the primary treatment in healthcare contexts. W represents an intermediate outcome, while B, given based on values of L, A, W, would represent salvage therapy or second line treatment in cases of poor response to A. Y represents the final outcome of interest.

For a single treatment A, *the effect of treatment on the treated* (ETT) is defined as $\mathbb{E}[Y(a)|A = a] - \mathbb{E}[Y(a')|A = a]$. Such an effect can be viewed as a version of the ACE among the set of people naturally exposed to a particular level of the treatment. For example, ETT compares the effect of smoking one pack of cigarettes a day to smoking nothing in the set of people who happen to smoke one pack of cigarettes a day. In the causal model represented by Fig. 1 (a), the ETT is identified as $\mathbb{E}[Y|a] - \sum_C \mathbb{E}[Y|a', C]p(C|a)$.

For multiple treatments, the effect of treatment on the multiply treated is defined similarly. In Fig. 1 (b), this effect is defined as $\mathbb{E}[Y(a, b)|a, b] - \mathbb{E}[Y(a', b')|a, b]$. While it can be shown that in Fig. 1 (b) the ACE $\mathbb{E}[Y(a, b)] - \mathbb{E}[Y(a', b')]$ is identified as $\sum_{L,W} \mathbb{E}[Y|W, L, a, b]p(W|b, L)p(L) - \mathbb{E}[Y|W, L, a', b']p(W|a', L)p(L)$, the ETT is not identified. This is due to the fact that the second term of the ETT is a function of variables $Y(a', b')$ and B, where the former is defined via (1) as $Y(a', b', W(a', L), L)$, while the latter is defined via (1) as $B(W(A, L), A(L), L)$. In other words, the ETT is a function of a joint distribution containing $p(W(a'), W)$ as a marginal, which is not identified under the structural causal model. This issue is described in detail in (Shpitser and Tchetgen Tchetgen, 2016).

The ACE and the ETT, where variables are manipulated to constants in order to mimic RCTs, are contrasts of substantive interest in applied settings such as econometrics and public health. In settings such as precision medicine, variables are manipulated based on observed patient characteristics, with the aim of improving positive outcomes or minimizing harmful ones. The resulting counterfactuals are defined as follows. For every $A_i \in A$ to be manipulated, define a set $L_i \subseteq V \setminus A$ to be some set of variables not causally determined by A_i (graphically, this means there is no directed path from A_i to any element in L_i in $\mathcal{G}$). Given a set of functions $f \equiv \{f_i : \mathfrak{X}_{L_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$, we define the response $Y(f)$ to setting values of each $A_i \in A$ according to its corresponding function f_i as

$$Y(\{A_i{=}f_i(L_i(f)){:}A_i{\in}\,\mathrm{pa}(Y){\cap}A\},\{W(f){:}W{\in}\,\mathrm{pa}(Y){\setminus}A\}),$$

by analogy with (1). As an example, in the model shown in Fig. 1 (a), given a function $f_A : \mathfrak{X}_L \mapsto \mathfrak{X}_A$, $Y(f_A) \equiv Y(A = f_A(L), L)$. Similarly, in the model shown in Fig. 1 (b), given functions $f_A : \mathfrak{X}_L \mapsto \mathfrak{X}_A$, and $f_B : \mathfrak{X}_{L,W} \mapsto \mathfrak{X}_B$, $Y(f_A, f_B) \equiv Y(L, A = f_A(L), W(L, A = f_A(L)), B = f_B(L, W(L, A = f_A(L))))$. Here the second response of Y is defined according to value of B set by f_B using values of W recursively determined by counterfactually setting A according to f_A. Functions in the set of f above are also known as dynamic treatment regimes (DTRs).

As before, if all variables in a causal model are observed, $p(V(f))$ is identified for any set $A \subseteq V$, and set of functions $f_A \equiv \{f_i : A_i \in A\}$ by the following variation of (2):

$$\prod_{V_i \in V} p(V_i | \{A_i = f_i(L_i) : A_i \in \mathrm{pa}(V_i) \cap A\}, \mathrm{pa}(V_i) \setminus A).$$

Responses of specific variables in V to A being set according to f is obtained from the above formula by marginalization, as before. As an example, $p(Y(f_A)) = \sum_L p(Y|A = f_A(L), L)p(L)$ in Fig. 1 (a).

Having described the ETT and responses to DTRs, we are now ready to describe shift interventions. Assume we are interested, in Fig. 1 (a), in the outcome Y had the drug dose A been changed from its given value a by a known function f_A. We define such a counterfactual, by analogy with (1) as $Y(A = f_A(A), L)$. Note that unlike $Y(a)$, each person in the data is assigned a potentially different dose, as would be the case for responses to DTRs. However, unlike DTR counterfactuals, the function only uses values of A as inputs.

Assuming A, B in Fig. 1 (b) also represent drug doses administered over time, we may be interested in how the outcome Y changes had drug doses been changed from their values by known functions f_A, f_B. Note that there are two ways to define such a counterfactual, which diverge in how the second treatment A_2 is manipulated.

One definition might consider the response of Y to the first treatment A being given by a fixed function f_A *of the observed treatment* A, and the second treatment B being given by a fixed function f_B *of the observed treatment* B. This response $Y(A = f_A(A), B = f_B(B))$ is defined as $Y(A = f_A(A), B = f_B(B), W(A = f_A(A), L), L)$. Another definition might consider the response of Y to the first treatment A being given by a fixed function f_A *of the observed treatment* A, and the second treatment B being given by a fixed function f_B *of the treatment* B *observed in the world where the first treatment* A *was counterfactually shifted according to* f_A. This response $Y(A = f_A(A), B = f_B(B(A = f_A(A))))$ is defined as $Y(L, A = f_A(A), W(L, A = f_A(A)), B = f_B(\tilde{B}))$, where $\tilde{B} \equiv B(L, A{=}f_A(A), W(L, A{=}f_A(A)))$.

We call the first definition *shift interventions on the treated* (SITs), and the second definition *shift interventions as policies* (SIPs). Unsurprisingly, identification theory for SITs bears some similarity to that of ETTs, while identification theory for SIPs bears some similarity to that of DTRs, although in both cases new subtleties present themselves.

SITs are of interest whenever deviations from current best practices are investigated. For instance, responses to SITs would be the correct counterfactual to use in healthcare settings to investigate the effect of dosing changes from an existing standard. SIPs are of interest when variable manipulations have a compound effect, and therefore effects of prior shift interventions on intermediate outcomes must be taken into account. For instance, responses to SIPs could be used to evaluate changes to financial aid, or a medical treatment administered over time with a compound effect. SIPs have been described, under a different name, in section 5.1 in (Richardson and Robins, 2013).

Before describing identification theory for SITs and SIPs, we give their general definitions, using a modification of (1). Fix $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$. By analogy with (1), we define for any $Y \in V$, the counterfactual response $Y(f(A))$ to SITs on A as

$$Y(\{\tilde{A}{=}\tilde{f}(\tilde{A}) : \tilde{A} \in A \cap \mathrm{pa}(Y)\}, \{W(f(A)) : W \in \mathrm{pa}(Y) \setminus A\}),$$

and the counterfactual response $Y(f)$ to SIPs on A as

$$Y(\{\tilde{A}{=}\tilde{f}(\tilde{A}(f)) : \tilde{A} \in A \cap \mathrm{pa}(Y)\}, \{W(f) : W \in \mathrm{pa}(Y) \setminus A\}).$$

3 IDENTIFICATION UNDER FULL OBSERVABILITY

We first describe identification theory for SITs and SIPs in cases where all variables in a causal model are observed. Identification for SIPs in fully observed models is given by the following result.

Theorem 1 *Fix* $A \subseteq V$*, and a set of functions* $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$ *in a fully observed functional causal model given by the DAG* $\mathcal{G}$*. Then* $p(V(f))$ *is identified and equal to*

$$\prod_{V_i \in V} p(V_i \mid \{A_i = f_i(A_i) : A_i \in \mathrm{pa}(V_i)\}, \mathrm{pa}(V_i) \setminus A_i).$$

For example, given f_A, f_B in Fig. 1 (b), $p(\{L, A, W, B, Y\}(f_A, f_B))$ is identified as $p(L)p(A|L)p(W|A = f_A(A), L)p(B|W, A = f_A(A), L)\ p(Y|B = f_B(B), W, A = f_A(A), L)$, and so $p(Y(f_A, f_B))$ is equal to

$$\sum_{L,W,A,B} p(Y|B{=}f_B(B),W,A{=}f_A(A),L)p(L)p(A|L)$$
$$p(B|W, A = f_A(A), L)p(W|A{=}f_A(A),L).$$

That is, identification of responses to SIPs in fully observed models resembles identification of DTRs.

Now let us consider identification of responses to SITs. It turns out that even if the causal model is fully observed, SITs may not be identified if multiple treatments are manipulated simultaneously, due to the same issue that prevents identification of ETTs. We have the following result.

Theorem 2 *Fix disjoint* $A, Y \subseteq V$*, and a set of* unrestricted *functions* $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$ *in a fully observed functional causal model given by the DAG* $\mathcal{G}$*.*

Fix the set of all directed paths π *in* $\mathcal{G}$ *which start with* $A_i \in A$*, end in some element in* $A \cup Y$*, and which do not intersect elements in* $A \cup Y$ *otherwise. Then* $p(Y(f(A)))$ *is identified if and only if there are no two elements in* π *which share the first edge and where one path ends in an element in* A*, and another path ends in an element in* Y*. Moreover, if* $p(Y(f(A)))$ *is identified, it is equal to*

$$\sum_{Y^* \setminus Y} \prod_{V_i \in Y^* \setminus \tilde{Y}} p(V_i | \mathrm{pa}(V_i)) \times$$
$$\prod_{V_i \in \tilde{Y}} p(V_i | \{A_i = f_i(A_i) : A_i \in A \cap \mathrm{pa}(V_i)\}, \mathrm{pa}(V_i) \setminus A),$$

where Y^* *is the set of ancestors of* Y *in* $\mathcal{G}$*, and* $\tilde{Y}$ *is the set of variables not in* A *which lie on a path in* π *that ends in* Y*.*

For example, given f_A, f_B, $p(Y(f_A(A), f_B(B)))$ is not identified in Fig. 1 (b), since the set of directed paths in π will contain $B \to Y$, $A \to Y$, $A \to W \to Y$, $A \to W \to B$, and $A \to B$. Since $A \to W \to Y$ and $A \to W \to B$ share the first edge, and have final elements in Y and B, the condition of theorem 2 applies.

However, if we consider identification of the same distribution $p(Y(f_A(A), f_B(B)))$ in Fig. 1 (c), where the edge $W \to B$ is absent, we obtain identification:

$$\sum_{L,A,W,B} (p(B|W,A,L)p(A|L)p(L)) \times \quad (4)$$
$$(p(Y|B{=}f_B(B),W,A{=}f_A(A),L)p(W|A{=}f_A(A),L))$$

Note that while identification of ETTs and SITs in fully observed DAGs runs into a similar difficulty having to do with *recanting witnesses* (Avin et al., 2005), identification results for these two types of counterfactuals are nevertheless quite different. This is because ETTs are defined as functions of *counterfactual conditionals* $p(Y(a)|A = a')$ for some set A, while SITs are defined as *counterfactual marginals*.

4 IDENTIFICATION WITH HIDDEN VARIABLES

Most causal inference problems of practical importance contain hidden but relevant variables, motivating the use of causal models of a DAG where some variables are not observed. As we now show, identification theory implied by the structural causal model of DAGs with hidden variables is more involved for both SIPs and SITs.

Identification theory of a causal model of a DAG $\mathcal{G}$ with vertices $V \cup H$, where V corresponds to observed variables and H corresponds to hidden variables is often phrased on an acyclic directed mixed graph (ADMG) called a latent projection (Verma and Pearl, 1990). By an ADMG we mean a graph with directed ($\to$) and bidirected ($\leftrightarrow$) edges and no directed cycles.

Given a DAG $\mathcal{G}(V \cup H)$ where V are observed variables and H are hidden variables, we define the *latent projection* ADMG $\mathcal{G}(V)$ with vertices V as follows. For every $V_i, V_j \in V$, if there exists in $\mathcal{G}$ a directed path from V_i to V_j with all intermediate vertices in H, an edge $V_i \to V_j$ exists in $\mathcal{G}(V)$. For every V_i, V_j, if there exists a collider-free path from V_i to V_j in $\mathcal{G}$ with the first edge on the path of the form $V_i \leftarrow$ and the last edge on the path of the form $\to V_j$, an edge $V_i \leftrightarrow V_j$ exists in $\mathcal{G}(V)$. For example, if L is unobserved in Fig. 1 (a), then the resulting latent projection is shown in Fig. 1 (e). This example illustrates that latent projections are not always simple graphs.

Latent projections are used because for any two distinct DAGs $\mathcal{G}_1(V \cup H_1)$, $\mathcal{G}_2(V \cup H_2)$ that share the same latent projection $\mathcal{G}(V) \equiv \mathcal{G}_1(V) = \mathcal{G}_2(V)$ also share non-parametric identification theory (Richardson et al., 2017).

Before describing this theory, we introduce a few additional definitions we will need. Given an ADMG $\mathcal{G}$, and $S \subseteq V$, define the induced subgraph $\mathcal{G}_S$ to be a graph containing vertices in S, and any edge in $\mathcal{G}$ connecting elements of S. Given an ADMG $\mathcal{G}$, a district of $\mathcal{G}$ is a bidirected-connected component. The set of districts of $\mathcal{G}$ forms a partition of vertices in $\mathcal{G}$, and is denoted by $\mathcal{D}(\mathcal{G})$. Finally, given a set S in $\mathcal{G}$, define $\mathrm{pa}(S) \equiv \bigcup_{S_i \in S} \mathrm{pa}(S_i)$.

Identification theory in hidden variable models uses ADMGs in an analogous way identification theory in fully observed models uses DAGs. Just as the structural causal model defined on a fully observed DAG $\mathcal{G}(V)$ implies the DAG factorization on the observed data distribution with respect to $\mathcal{G}(V)$, and identification of *all* interventional distributions $p(V(a))$ in terms of a modified factorization of $\mathcal{G}$, so does the structural causal model defined on a hidden variable DAG $\mathcal{G}(V \cup H)$ implies the *nested Markov factorization* (Richardson et al., 2017) on the observed data distribution with respect to the latent projection ADMG $\mathcal{G}(V)$, and identification of certain marginal interventional distributions $p(Y(a))$ in terms of a modified nested factorization of $\mathcal{G}(V)$ given by the *ID algorithm* (Tian and Pearl, 2002; Shpitser and Pearl, 2006).

The nested Markov factorization of $p(V)$ with respect to an ADMG $\mathcal{G}(V)$ is defined in terms of *Markov kernels* of the form $q_S(S \mid W_S)$, with a single kernel for each subset $S \subseteq V$ that is *an intrinsic set*. A Markov kernel $q_S(S \mid W_S)$ is any map from $\mathfrak{X}_{W_S}$ to normalized densities over S. For any $A \subseteq S$, conditioning and marginalization in Markov kernels is defined in the usual way as:

$$q_S(A|W_S) \equiv \sum_{S \setminus A} q_S(S|W_S);\ q_S(S|A,W_S) \equiv \frac{q_S(S|W_S)}{q_S(A|W_S)}.$$

A set S is intrinsic in $\mathcal{G}$ if $\mathcal{G}_S$ contains a single district and is reachable in $\mathcal{G}$. A set S is said to be reachable in $\mathcal{G}$ if there exists a sequence of ADMGs $\mathcal{G}_1, \ldots, \mathcal{G}_k$ such that $\mathcal{G}_1 \equiv \mathcal{G}$, $\mathcal{G}_k \equiv \mathcal{G}_S$, each $\mathcal{G}_i$ is obtained from $\mathcal{G}_{i+1}$ by removing a specific vertex V_i and all edges with V_i as one endpoint. Finally, for each $\mathcal{G}_{i+1}$, the vertex V_i to be removed to obtain $\mathcal{G}_i$ has no directed and bidirected (consisting entirely of $\leftrightarrow$ edges) path to *any* other vertex V_j in $\mathcal{G}_{i+1}$.

The Markov kernels defining the nested Markov models are always functionals of $p(V)$. For example, in Fig. 1 (d), the Markov kernels corresponding to all intrinsic sets are:

$$q_A(A) = p(A);\ q_M(M|A) = p(M|A);$$
$$q_{\{Y,A\}}(Y,A|M) = p(Y|A,M)p(M);$$
$$q_Y(Y|M) = \sum_A p(Y|M,A)p(A).$$

We describe the general scheme for deriving functionals for intrinsic Markov kernels from $p(V)$ in the Supplement.

The nested Markov factorization expresses $p(V)$ and any kernel $q_R(R \mid W_R)$ where R is a reachable set in terms of Markov kernels corresponding to intrinsic sets, as follows:

$$p(V) = \prod_{D \in \mathcal{D}(\mathcal{G}(V))} q_D(D \mid W_D),$$
$$q_R(R \mid W_R) = \prod_{D \in \mathcal{D}(\mathcal{G}(V)_R)} q_D(D \mid W_D).$$

For instance, the nested Markov factorization for the ADMG in Fig. 1 (d) implies $p(Y, M, A) = q_{\{Y,A\}}(Y, A|M)q_M(M|A)$, which is sometimes called the district or c-component factorization of an ADMG.

Given disjoint subsets Y, A of V, the nested Markov factorization naturally leads to the following reformulation of the complete algorithm for identification of $p(Y(a))$, sometimes called the *ID algorithm* (Shpitser and Pearl, 2006). This algorithm can be expressed as a modified nested Markov factorization as follows:

$$p(Y(a)) = \sum_{Y^* \setminus Y} \prod_{D \in \mathcal{G}(V)_{Y^*}} q_D(D \mid W_D)|_{\{A_i = a_i : A_i \in W_D \cap A\}},$$

where Y^* is ancestors of Y in $\mathcal{G}_{V \setminus A}$. This factorization is defined provided each D on the right hand side is intrinsic, otherwise it is undefined and $p(Y(a))$ is not identified given the structural causal model for any hidden variable DAG $\mathcal{G}(V \cup H)$ that yields the latent projection $\mathcal{G}(V)$.

For example, in the graph shown in Fig. 1 (d), we have:

$$p(Y(a)) = \sum_M \underbrace{\left(\sum_A p(Y|M, A)p(A) \right)}_{q_Y(Y|M)} \underbrace{p(M|a)}_{q_M(M|A=a)},$$

known as the *front-door formula*, while $p(Y(a))$ is not identified in Fig. 1 (e).

Identification of SIPs can be characterized in terms of the nested Markov factorization, with an additional subtlety, by the following result.

Theorem 3 *Fix disjoint subsets $A, Y \subseteq V$, and a set of* unrestricted *functions $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$ in a functional causal model given by the DAG $\mathcal{G}(V \cup H)$ that yields the latent projection ADMG $\mathcal{G}(V)$. Define Y^* as the set of ancestors of Y in $\mathcal{G}(V)$. Then $p(Y(f))$ is identified if and only if for some district $D \in \mathcal{D}(\mathcal{G}_{Y^*})$, no element of A in D has children in D in $\mathcal{G}_D$. Moreover, if $p(Y(f))$ is identified, it is equal to*

$$\sum_{Y^* \setminus Y} \prod_{D \in \mathcal{D}(\mathcal{G}_{Y^*})} q_D(D|W_D)|_{\{A_i = f_i(A_i) : A_i \in A \cap \mathrm{pa}(D)\}}$$

As an example, the distribution $p(Y(f))$ in Fig. 1 (d) is identified, since the districts of ancestors of Y are $\{A, Y\}$, and $\{M\}$, and no district contains a child of A in the induced subgraph for that district. The identifying formula is $\sum_{M,A} p(Y|A, M)p(A)p(M|A = f(A))$. On the other hand, the distribution $p(Y(f))$ in Fig. 1 (e) is not identified, even though the single district among the ancestors of Y, namely $\{A, Y\}$, is intrinsic. This is because this district contains a child of A.

Identification of SITs is a little more involved, as we must also ensure the difficulty described with the ETT, where the counterfactual is a function of a non-identified marginal of the form $p(W(A_i = f_i(A_i)), W)$ is avoided.

Theorem 4 *Fix disjoint subsets $A, Y \subseteq V$, and a set of* unrestricted *functions $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$ in a functional causal model given by the DAG $\mathcal{G}(V \cup H)$ that yields the latent projection ADMG $\mathcal{G}(V)$. Fix the set of all directed paths π in $\mathcal{G}(V)$ which start with $A_i \in A$, end in some element in $A \cup Y$, and which do not intersect elements in $A \cup Y$ otherwise. Define Y^* as the set of ancestors of Y in $\mathcal{G}(V)$. Then $p(Y(f(A)))$ is identified if and only if*

- *There are no two paths in π which start with the same edge, and where one path ends in an element of Y, and another in an element of A.*
- *Every element of A that lies in a district D in $\mathcal{G}(V)_{Y^*}$ does not have children in D in $\mathcal{G}_D$.*
- *For any two paths in π where the second vertex on the path is in district D, either both paths have the final element in A or both paths have the final element in Y.*

Moreover, if $p(Y(f(A)))$ is identified, it is equal to

$$\sum_{Y^* \setminus Y} \prod_{D \in \mathcal{D}(\mathcal{G}_{Y^*})} q_D(D|W_D)|_{\{A_i = f_i(A_i) : A_i \in A \cap \mathrm{pa}^Y(D)\}},$$

where $\mathrm{pa}^Y(D)$ are parents of D along edges that are first edges on paths in π that end in Y.

As an example of the application of this theorem, consider Fig. 1 (f), where we are interested in identifying $p(Y(A_1 = f_1(A_1), A_2 = f_2(A_2)))$. If all green edges are absent, the conditions of the theorem are satisfied, and this distribution is identified, in fact by the same functional as in (4). If the edge (1) is present, identification fails because of the presence of paths $A \to W \to B$ and $A \to W \to Y$, as in Theorem 2. If the edge (2) is present, there exists a district in Y^*, namely $\{A, L, W, Y\}$ with an element A in the district that also has a child in the district (W). If the edge (3) is present, a path $A \to B$ ends in a treatment, while a path $A \to Y$ ends in an outcome, and both paths have a second vertex in the same district.

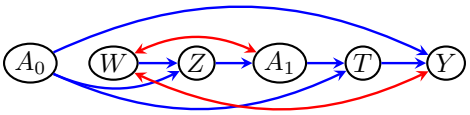


Figure 2: An example where SITs and SIPs give different identifying functionals for $p(Y(f))$ and $p(Y(f(A)))$ respectively.

A Note On Completeness

Completeness results in this section have specified an *unrestricted* set of shift functions $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} \mid A_i \in A\}$, and only hold under a sufficiently large class of shift functions that allow counterexamples in our proofs to be constructed. Results of this type are in the spirit of non-parametric identification theory in the sense that shift functions act as a kind of user-specified structural equation, and identification theory results are often stated in a way that does not restrict structural equations. A similar notion of completeness for Tian's identification algorithm for responses to dynamic treatment regimes (Tian, 2008), was shown to hold in (Shpitser and Sherman, 2018).

Identification theory for sufficiently restricted classes of shift functions becomes considerably more complicated than stated here, and indeed it may be possible identification may be shown to hold even if the response to an unrestricted class of shift functions is not identified. The situation is similar to one where semi-parametric restrictions are placed on structural equations in a causal model.

It is also worth noting we will always have identification when shift functions are specified as identity functions, in which case the interventional distributions of $p(Y(f))$ or $p(Y(f(A)))$ are equal to $p(Y)$.

Differences In Identifying Functionals

We now give another example that illustrates that when SITs and SIPs that involve multiple treatments are identified, they will in general give different identifying functionals. Consider the hidden variable causal model represented by a graph in Fig. 2, where Y is the outcome of interest, and we are interested in its response to both SITs and SIPs on treatment variables A_0 and A_1.

Here, $Y^* = \{Y, T, A_1, Z, W, A_0\}$, and the set of districts in $\mathcal{D}(\mathcal{G}_{Y^*})$ are $D_1 = \{A_0\}, D_2 = \{W, A_1, Y\}, D_3 = \{Z\}, D_4 = \{T\}$. We first note that the SIP $p(Y(f))$ is identified because no elements of A in some district D have children in that district. In particular, for $A_0 \in D_1 = \{A_0\}, \mathrm{ch}_{\mathcal{G}_{D_1}}(A_0) = \emptyset$, and for $A_1 \in D_2 = \{W, A_1, Z, T\}$.

The corresponding sets $\mathrm{pa}(D)$ for each district D are $\mathrm{pa}(D_1) = \emptyset, \mathrm{pa}(D_2) = \{A_0, Z, T\}, \mathrm{pa}(D_3) = \{W, A_0\}, \mathrm{pa}(D_4) = \{A_0, A_1\}$, and therefore $A \cap \mathrm{pa}(D_1) = \emptyset, A \cap \mathrm{pa}(D_2) = \{A_0\}, A \cap \mathrm{pa}(D_3) = \{A_0\}, A \cap \mathrm{pa}(D_4) = \{A_0, A_1\}$.

The identifying functional from applying theorem 3 is therefore

$$\sum_{A_0,W,Z,A_1,T} \{p(A_0)\}\{p(Y|A_0 = f_0(A_0), Z, W, A_1, T) \times p(W)p(A_1|A_0 = f_0(A_0), Z, W)\} \times \{p(Z|W, A_0 = f_0(A_0))\} \times \{p(T|A_0 = f_0(A_0), A_1 = f_1(A_1))\},$$

with each term corresponding to the districts in Y^* enclosed in braces.

Next, consider the SIT $p(Y(f(A)))$ for $A = \{A_0, A_1\}$. We note that $A \cup Y = \{A_0, A_1, Y\}$. Y^* is unchanged, and all three identification conditions are satisfied. The set of paths π are

$$\pi = \{(A_0, T, Y), (A_0, Z, A_1), (A_1, T, Y), (A_0, Y)\}$$

$\mathrm{pa}^Y(D)$ for each district are $\mathrm{pa}^Y(D_1) = \emptyset, \mathrm{pa}^Y(D_2) = \{A_0\}, \mathrm{pa}^Y(D_3) = \emptyset, \mathrm{pa}^Y(D_4) = \{A_1\}$. $\mathrm{pa}^Y(D_2)$ only includes A_0, as there is only one path ending in Y whose first edges are parents of D_2 – namely, $A_0 \to Y$. $\mathrm{pa}^Y(D_3)$ is empty since no such paths exist. $A \cap \mathrm{pa}^Y(D)$ for each D gives $A \cap \mathrm{pa}^Y(D_1) = \emptyset, A \cap \mathrm{pa}^Y(D_2) = \{A_0\}, A \cap \mathrm{pa}^Y(D_3) = \emptyset, A \cap \mathrm{pa}^Y(D_4) = \{A_1\}$,

which means that the identifying functional is changed in exactly one place – $p(Z|W, A_0 = f_0(A_0))$ is replaced with $p(Z|W, A_0)$, yielding:

$$\sum_{A_0,W,Z,A_1,T} \{p(A_0)\}\{p(Y|A_0 = f_0(A_0), Z, W, A_1, T) \times p(W)p(A_1|A_0 = f_0(A_0), Z, W)\} \times \{p(Z|W, A_0)\} \times \{p(T|A_0 = f_0(A_0), A_1 = f_1(A_1)\}.$$

Once again, each term corresponding to the districts in Y^* is enclosed in braces.

5 PARAMETRIC AND SEMI-PARAMETRIC INFERENCE

Assessing the impact of responses to SIPs and SITs entails evaluating functions of counterfactual distributions $p(Y(f))$ and $p(Y(f(A)))$ from data. Here we concentrate on estimating expected value parameteres β in cases where these distributions are identified, e.g. $\mathbb{E}[Y(f)]$, and $\mathbb{E}[Y(f(A))]$.

If a parametric model for the observed data distribution $p(V)$, or a sufficiently large part of the distribution, can

be correctly specified, maximum likelihood plug-in estimators are used for efficient statistical inference for β. In the fully observed model, plug-in estimators may be straightforwardly derived from a DAG observed data likelihood. For example, $\mathbb{E}[Y(f_A(A), f_B(B))]$ with respect to the distribution in (4) may be estimated via

$$\frac{1}{n}\sum_i \sum_{B,W} p(B|W, A_i, L_i; \hat{\eta}_B)p(W|A=f_A(A_i), L_i; \hat{\eta}_W)$$
$$\mathbb{E}[Y|B=f_B(B), W, A=f_A(A_i), L_i; \hat{\eta}_Y],$$

where $\hat{\eta}_B, \hat{\eta}_W, \hat{\eta}_Y$ are maximum likelihood estimates of parameters for parametric models above.

If β is identified in a hidden variable model with a latent projection ADMG $\mathcal{G}(V)$, parametric statistical inference is sometimes possible using plug-in estimators that maximize nested Markov likelihoods, which are known for discrete data (Evans and Richardson, 2018), and multivariate normal distributions (Shpitser et al., 2018). We do not discuss these estimators further in the interests of space.

If a parametric likelihood cannot be assumed, statistical inference must proceed within a semi-parametric or non-parametric model, where a part of the likelihood or the whole likelihood is infinite-dimensional. In such cases, plug-in estimators are known to have non-negligible first order bias. A principled alternative approach to obtaining high quality consistent estimators is based on the semi-parametric theory, and influence functions (Tsiatis, 2006).

The resulting *regular asymptotically linear* (RAL) estimators take the form

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi(Z_i) + o_p(1),$$

where $\phi \in \mathbb{R}^q$ with mean zero and finite variance, $o_p(1)$ denotes a term that approaches to zero in probability, and $\phi(Z_i)$ is the *influence function (IF)* of the ith observation for the parameter vector β. RAL estimators are consistent and asymptotically normal (CAN), with the variance of the estimator given by its IF:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \phi\phi^T).$$

Thus, there is a bijective correspondence between RAL estimators and IFs.

We now derive the IF for β in a single treatment setting given by Fig. 1 (a), where SITs and SIPs coincide.

Theorem 5 *Fix $\beta = \sum_{C,A} \mathbb{E}[Y|a=f(A), C]p(A|C)p(C)$, which is equal to $\mathbb{E}[Y(f(A))] = \mathbb{E}[Y(f)]$ under the model in Fig. 1 (a). The efficient influence function for β under the non-parametric observed data model is given by*

$$U(\beta) = \frac{\sum_{A'} \mathbb{I}(A = f(A'))p(A' \mid C)}{p(A \mid C)}\{Y - \mathbb{E}[Y \mid A, C]\}$$
$$+ \mathbb{E}[Y \mid a = f(A), C)] - \beta \quad (5)$$

The influence function $U(\beta)$ leads to a RAL estimator which solves the estimating equation $\mathbb{E}[U(\beta)] = 0$, and which resembles augmented inverse probability weighted (AIPW) estimators derived in other contexts in causal inference (Scharfstein et al., 1999). As is often the case with these estimators, our estimator exhibits the property of *double robustness*, where the estimator remains consistent in the union model where either $\mathbb{E}[Y|A, C]$ or $p(A|C)$ is correctly specified.

Theorem 6 *The estimator of β which solves the estimating equation $\mathbb{E}[U(\beta)] = 0$ is consistent, and asymptotically normal (CAN) in the union model where one of $\pi(C; \eta_A) = p(A|C)$, $m(A, C; \eta_Y) = \mathbb{E}[Y|A, C]$ is correctly specified.*

In the Supplement we also derive the efficient influence function for the shift intervention p(Y(f)) in a variant of the causal model shown in Fig. 1 (d) that also contains a vector of baseline covariates.

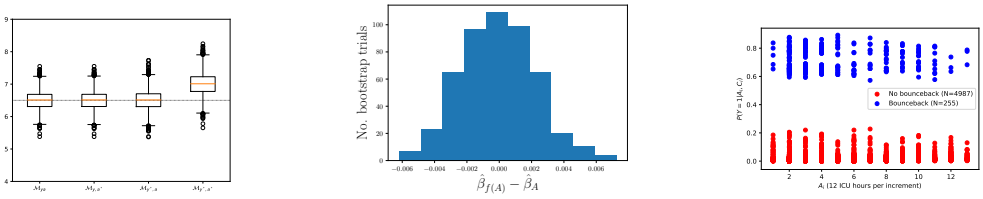
6 SIMULATIONS AND A DATA APPLICATION

We now present a simulation study that demonstrates our estimator is doubly robust to misspecification of either the $E[Y|A, C]$ model or the $p(A|C)$ model. The precise data generating process is described in the Supplement.

Based on the simulation above, our parameter of interest $\beta = E[Y(f(A))]$, where $f(A) = A + 0.5$, is equal to 6.5. We simulated datasets of size 500 and used 5000 replicates. The results are seen in Fig. 3a, where $\mathcal{M}_{y,a}$ denotes the correctly specified models for $\mathbb{E}[Y|A, C]$, and $p(A|C)$, $\mathcal{M}_{y,a^*}$ denotes the model where only $\mathbb{E}[Y|A, C]$ is specified correctly, $\mathcal{M}_{y^*,a}$ denotes the model where only $p(A|C)$ is specified correctly, and $\mathcal{M}_{y^*,a^*}$ denotes the model where both $\mathbb{E}[Y|A, C]$ and $p(A|C)$ are specified incorrectly. As expected, the estimates show no bias for $\mathcal{M}_{y,a}$, $\mathcal{M}_{y,a^*}$, and $\mathcal{M}_{y^*,a}$, while bias is introduced in the model $\mathcal{M}_{y^*,a^*}$.

Data Application

We now describe our data application. Intensive care unit (ICU) readmission ("bounceback") after cardiac surgery is costly and associated with worse mortality and morbidity outcomes (Benetis et al., 2013). We used our



(a) (b) (c)

Figure 3: (a) Estimation of $\mathbb{E}[Y(f(A))]$ using (6) under various types of model misspecification. (b) Empirical distribution ($N = 500$) of $\hat{\beta}_{f(A)} - \hat{\beta}_A$, with 95% confidence interval (-0.0043, 0.0047). (c) The bounceback probability (Y axis) learned by the random forest model for $\mathbb{E}[Y|A, C]$ vs discretized length of stay (X axis) for all patients in the data set. Blue values denote bounceback actually occurred, red values indicate bounceback did not actually occur.

methods to estimate shift interventions to investigate whether increasing length of stay may influence the probability of bounceback. Data from 5242 patient visits to our institution who had undergone a surgical procedure on the heart, entered the hospital ICU at any point, and did not die during the visit were curated from our institution's contribution to the Society of Thoracic Surgeon Adult Cardiac Surgery database, and our internal electronic health records. 151 discrete and continuous variables covering patient demographics, medications, as well as pre-, inter- and post-operative status were used.

We partitioned variables in the dataset into three types: the treatment variable A which is the number of initial ICU hours, discretized into 12-hour time intervals, the binary outcome Y representing bounceback, and a vector C of covariates representing potential confounders. We discretized A to avoid issues with lack of support. Specifically, we avoid unstable or invalid inferences which occur if $p(A|C) = 0$. We are interested in the change in probability of bounceback after a hypothetical increase of length of stay by 24 hours. We estimate this probability by using (6), where the outer expectation is evaluated empirically, and the required nuisance models $p(A|C)$ and $\mathbb{E}[Y|A, C]$ are estimated via a negative binomial regression (in case of overdispersion) and a random forest classifier, respectively. We are interested in a policy where patients receive an additional 24 initial ICU hours, denoted $f(A) = A + 2$.

We compare the total effect under the shift intervention $\hat{\beta}_{f(A)} = \mathbb{E}[Y(f(A))]$ against the total effect under the observed distribution of A, $\hat{\beta}_A = \mathbb{E}[Y(A)] = \mathbb{E}[Y]$. The distribution for $\hat{\beta}_{f(A)} - \hat{\beta}_A$ under 500 bootstrap samples is given in Fig. 3b. As the 95% bootstrap confidence interval contains 0, we fail to reject the null of no statistically significant effect of the shift intervention of increased initial ICU hours on ICU readmission rates.

To explore why the null hypothesis was not rejected, we considered the behavior of the learned outcome regression function $\mathbb{E}[Y|A, C]$ with respect to A. Fig. 3c shows the predicted bounceback probabilities for each unit in our data, plotted vs their observed discretized length of stay. Red values denote no bounceback (the significantly more common case), while blue values denote bounceback. The response to the shift intervention that we estimated via (6) can be viewed as a modified empirical average of this regression, augmented with an inverse weighted term. The learned regression function appears to indicate that our data contains two types of patients: the significantly more common low risk patients, and the rarer high risk patients. Both types of patients occur at all durations of length of stay, and variations of length of stay are not a significantly predictive feature for type. In particular, variations in A do not significantly alter patient's risk from its level predicted from other features.

7 CONCLUSIONS

In this paper we define a type of soft intervention where a set of variables are manipulated to obtain values which are fixed functions of their previous values. We call this type of intervention *shift intervention*. We showed that if multiple variables are manipulated, shift interventions may be defined with respect to naturally occurring values of manipulated variables, or with respect to recursively defined values of manipulated variables responding to previous shift interventions. We gave a sound and complete identification algorithm for both types of shift interventions in fully observed and hidden variable causal models.

In addition, we derived an efficient semi-parametric estimator based on efficient influence functions for a special case of responses to shift interventions motivated by a clinical problem. We demonstrated the utility of our method by a simulation study, and applied it to consider how the readmission probability to the intensive care unit (ICU) of a hospital changes if the duration of the patients' stay in the ICU is manipulated to be longer.

References

C. Avin, I. Shpitser, and J. Pearl. Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, volume 19, pages 357–363. Morgan Kaufmann, San Francisco, 2005.

R. Benetis, E. Širvinskas, B. Kumpaitiene, and Š. Kinduris. A case-control study of readmission to the intensive care unit after cardiac surgery. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, 19:148–152, Feb. 2013. ISSN 1234-1010. doi: 10.12659/MSM.883814.

B. Chakraborty and E. Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. New York: Springer-Verlag, 2013.

F. Eberhardt. Direct causes and the trouble with soft interventions. *Erkenntnis*, 79(4):755–777, 2014.

R. J. Evans and T. S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 2018. (to appear).

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019. doi: 10.1080/01621459.2017.1422737.

D. Malinsky, I. Shpitser, and T. S. Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

J. Neyman. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in English. *Statistical Science*, 5:463–472, 1923.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo, 1988.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 978-0521895606.

T. S. Richardson and J. M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *preprint: http://www.csss.washington.edu/Papers/wp128.pdf*, 2013.

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs, 2017. Working paper.

J. M. Robins, S. D. Mark, and W. K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, pages 479–495, 1992.

D. B. Rubin. Causal inference and missing data (with discussion). *Biometrika*, 63:581–592, 1976.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1146, 1999.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto, 2006.

I. Shpitser and J. Pearl. Effects of treatment on the treated: identification and generalization. In *Uncertainty in Artificial Intelligence*, volume 25. AUAI Press, 2009.

I. Shpitser and E. Sherman. Identification of personalized effects associated with causal pathways. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.

I. Shpitser and E. J. Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433–2466, 2016.

I. Shpitser, R. J. Evans, and T. S. Richardson. Acyclic linear sems obey the nested markov property. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.

J. Tian. Identifying dynamic sequential plans. In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 554–561, Corvallis, Oregon, 2008. AUAI Press.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, pages 567–573, 2002. ISBN 0-262-51129-0.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag New York, 1st edition edition, 2006.

T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles, 1990.

8 Supplement

The supplement is organized as follows. Part A contains a description of the nested Markov model, and describes the functionals of the observed data distribution for every intrinsic Markov kernel that forms the factorization of this model. Part B contains details of our simulation study. Part C contains detailed proofs of all claims.

8.1 A. The Nested Markov Model, and Intrinsic Markov Kernels

We reproduce the standard definition of the nested Markov model found in Richardson et al. (2017). In particular, we show that every every Markov kernel $q_S(S|W_S)$ corresponding to an intrinsic set S in an ADMG $\mathcal{G}(V)$ is a functional of the observed data distribution $p(V)$ in the nested Markov model for $\mathcal{G}(V)$.

A conditional ADMG (CADMG) $\mathcal{G}(V, W)$ is an ADMG where a set of vertices in V are considered *random*, and a set of vertices in W are considered *fixed*. Any vertex $W_i \in W$ may not have edges with arrowheads into W_i in $\mathcal{G}(V, W)$. By convention, districts in a CADMG are defined with respect to random vertices only.

A variable $V_i \in V$ is said to be fixable in $\mathcal{G}(V, W)$ if no element $V_j \neq V_i$ in the district of V_i is a descendant of V_i. Given V_i fixable in $\mathcal{G}(V, W)$, define the graphical fixing operator $\phi_{V_i}(\mathcal{G}(V, W))$ that yields a new CADMG $\mathcal{G}(V \setminus \{V_i\}, W \cup \{V_i\})$ obtained from $\mathcal{G}(V, W)$ by changing the status of V_i from random to fixed, and removing all edges adjacent to V_i with arrowheads into V_i. We define *fixable* sequences of vertices in a CADMG $\mathcal{G}(V, W)$, as follows. The empty sequence $\langle\rangle$ is fixable in any graph. Given a non-empty sequence of the form $\sigma = \langle V_1, V_2, \ldots, V_k\rangle$, define the tail of the sequence $\tau(\sigma) \equiv \langle V_2, \ldots, V_k\rangle$. A sequence σ is fixable in $\mathcal{G}(V, W)$ if V_1 is fixable in $\mathcal{G}(V, W)$, and $\tau(\sigma)$ is a sequence fixable in $\phi_{V_1}(\mathcal{G}(V, W))$.

Given a CADMG $\mathcal{G}(V, W)$, a kernel $q_V(V|W)$, and V_i fixable in $\mathcal{G}(V, W)$, define the kernel fixing operator $\phi_{V_i}(q_V(V|W); \mathcal{G}(V, W))$ as:

$$\begin{aligned}\phi_{V_i}(q_V(V|W); \mathcal{G}(V, W)) &\equiv \frac{q_V(V|W)}{q_V(V_i|\,\mathrm{nd}(V_i), W)} \\ &= \tilde{q}_{V\setminus\{V_i\}}(V \setminus \{V_i\}|W \cup \{V_i\}),\end{aligned}$$

where $\mathrm{nd}(V_i)$ is the set of non-descendants of V_i in $\mathcal{G}(V, W)$, and the kernel in the denominator is obtained from $q_V(V|W)$ by marginalization and conditioning.

Given a fixable sequence $\sigma = \langle V_1, V_2, \ldots, V_k\rangle$ in $\mathcal{G}(V, W)$, and a kernel $q_V(V|W)$, define

$$\begin{aligned}\phi_{\langle\rangle}(\mathcal{G}(V, W)) &\equiv \mathcal{G}(V, W), \\ \phi_\sigma(\mathcal{G}(V, W)) &\equiv \phi_{\tau(\sigma)}(\phi_{V_1}(\mathcal{G}(V, W)).\end{aligned}$$

Similarly, define

$$\begin{aligned}\phi_{\langle\rangle}(q_V(V|W); \mathcal{G}(V, W)) &\equiv q_V(V|W), \\ \phi_\sigma(q_V(V|W); \mathcal{G}(V, W)) &\equiv \\ \phi_{\tau(\sigma)}(\phi_{V_1}(q_V(V|W); \mathcal{G}(V, W)); \phi_{V_1}(\mathcal{G}(V, W))).\end{aligned}$$

$p(V)$ is said to reside in the nested Markov model of the ADMG $\mathcal{G}(V)$ if for every fixable sequence σ, $\phi_\sigma(p(V); \mathcal{G}(V))$ obeys the *global Markov property* with respect to the CADMG $\phi_\sigma(\mathcal{G}(V, W))$, described in Richardson et al. (2017). If $p(V)$ is in the nested Markov model of $\mathcal{G}(V)$, then any two fixable sequences σ_1, σ_2 on a set W yield the same CADMG $\mathcal{G}(V \setminus W, W) = \phi_{\sigma_1}(\mathcal{G}(V)) = \phi_{\sigma_2}(\mathcal{G}(V))$, and the same kernel $q_{V\setminus W}(V \setminus W|W) = \phi_{\sigma_1}(p(V); \mathcal{G}(V)) = \phi_{\sigma_2}(p(V); \mathcal{G}(V))$. Thus we define, in the natural way, the fixing operators $\phi_S(\mathcal{G}(V))$, $\phi_S(p(V); \mathcal{G}(V))$ on sets $S \subseteq V$ if any fixable sequence exists for S in $\mathcal{G}(V)$.

If S is intrinsic in $\mathcal{G}(V)$, then the intrinsic Markov kernel $q_S(S|W_S)$ associated with the nested Markov model of $\mathcal{G}(V)$ is defined, in terms of $p(V)$, as $\phi_{V\setminus S}(p(V); \mathcal{G}(V))$, which is a functional of $p(V)$ by definition of $\phi()$.

8.2 B. The Data Generating Process for the Simulation Study

The data generating process for the simulation was a linear structural equation model. Specifically, the parameters and $f(A)$ were defined as

$$\begin{aligned}&C \sim N(0, 1), \\ &A = 4.5 + 2C + \epsilon_A, \\ &Y = 1.5 + 4C + A + \epsilon_Y, \\ &f(A) = A + 0.5,\end{aligned}$$

where both ϵ_A and ϵ_Y are drawn from standard normal distributions.

8.3 C. Proofs

Here we give proofs of all claims stated in the main body of the paper.

Theorem 1 *Fix $A \subseteq V$, and a set of functions $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$ in a fully observed functional causal model given by the DAG $\mathcal{G}$. Then $p(V(f))$ is identified and equal to*

$$\prod_{V_i \in V} p(V_i \mid \{A_i = f_i(A_i) : A_i \in \mathrm{pa}(V_i)\}, \mathrm{pa}(V_i) \setminus A_i).$$

Proof: The assumptions of the functional model (in fact a subset of assumptions encoding the weaker

FFRCISTG model Richardson and Robins (2013)) imply that for any $V_i \in V$, $V_i(a,b)$ is independent of $A(b)$ for any $B \in V \setminus \text{pa}(V_i)$ if $A = \text{pa}(V_i)$. We apply this independence restriction, along with the consistency property stating that $A(b) = a$ implies $V_i(a,b) = V_i(b)$, inductively to any term in the factorization of $p(V(f))$ of the form $p(V_i(\{A_i = f_i(A_i(f)) : A_i \in \text{pa}(V_i)\}, \{W(f) : \text{pa}(V_i) \setminus A_i)\})$ to yield our conclusion. See also a structurally similar proof of the soundness of the extended g-formula under the assumptions of the FFRCISTG model in Richardson and Robins (2013). □

Theorem 2 *Fix disjoint $A, Y \subseteq V$, and a set of* unrestricted *functions $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$ in a fully observed functional causal model given by the DAG $\mathcal{G}$.*

Fix the set of all directed paths π in $\mathcal{G}$ which start with $A_i \in A$, end in some element in $A \cup Y$, and which do not intersect elements in $A \cup Y$ otherwise. Then $p(Y(f(A)))$ is identified if and only if there are no two elements in π which share the first edge and where one path ends in an element in A, and another path ends in an element in Y. Moreover, if $p(Y(f(A)))$ is identified, it is equal to

$$\sum_{Y^* \setminus Y} \prod_{V_i \in Y^* \setminus \tilde{Y}} p(V_i | \text{pa}(V_i)) \times$$

$$\prod_{V_i \in \tilde{Y}} p(V_i | \{A_i = f_i(A_i) : A_i \in A \cap \text{pa}(V_i)\}, \text{pa}(V_i) \setminus A),$$

where Y^* is the set of ancestors of Y in $\mathcal{G}$, and $\tilde{Y}$ is the set of variables not in A which lie on a path in π that ends in Y.

Proof: Assume there exist two paths in π which share the first edge and where one path ends in an element in A, and another path ends in an element in Y. Consider the submodel of the causal model represented by ADMG $\mathcal{G}(V)$ where all bidirected edges are absent (in other words, in this submodel, unobserved confounders do not actually influence observed variables in any way, and it can be represented by a DAG $\mathcal{G}^\dagger$ which is an edge subgraph of the ADMG $\mathcal{G}(V)$ containing only $\rightarrow$ edges). Then in this submodel, either the preconditions of Lemma 4.2 in Shpitser and Tchetgen Tchetgen (2016) hold, or $p(Y(a), A)$, for any fixed assignment a given by $f(A)$, is expressible as a path-intervention, but cannot be rephrased as an edge intervention. A generalization of (1) that defines path and edge interventions is given as (3) and (4), respectively in Shpitser and Tchetgen Tchetgen (2016).

If $p(Y(a), A)$ is expressible as a path intervention, but not an edge intervention, then by Theorem 5.2 in Shpitser and Tchetgen Tchetgen (2016), $p(Y(a), A)$ is not identifiable. Regardless of whether the preconditions of Lemma 4.2 in Shpitser and Tchetgen Tchetgen (2016) hold, or the preconditions of Theorem 5.2 in Shpitser and Tchetgen Tchetgen (2016) hold, the non-identification is established for $p(Y_i(a_i), A_j)$, for a specific $A_i, A_j \in A$, $Y_i \in Y$, in a subgraph $\mathcal{G}^*$ of $\mathcal{G}^\dagger$ containing two (possibly overlapping) directed paths from A_i to A_j and A_i to Y_i. By definition of SIPs, A_j must be ancestral of some $Y_j \in Y$, via a directed path $A_j \rightarrow W_1 \rightarrow \ldots \rightarrow W_k \rightarrow Y_j$.

Since f are unrestricted, we consider f_i to be a function that simply sets A_i to a_i and ignores natural values of A_i, and f_j to be a function that sets A_j to the value A_j assumes naturally. This immediately implies that $p(Y_j(f_i, f_j), Y_i(f_i, f_j))$ is not identified in $\mathcal{G}^*$ by the proof of Theorem 5.2, which implies $p(Y(f))$ is not identified in $\mathcal{G}^\dagger$ and thus also in $\mathcal{G}$.

If no two paths in π exist with the given properties, then $p(Y(a), A)$ is expressible as an edge intervention for any value a given by f. The result then follows by definition of π, Y^* and $\tilde{Y}$, and Theorem 5.2 in Shpitser and Tchetgen Tchetgen (2016). □

Theorem 3 *Fix disjoint subsets $A, Y \subseteq V$, and a set of* unrestricted *functions $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$ in a functional causal model given by the DAG $\mathcal{G}(V \cup H)$ that yields the latent projection ADMG $\mathcal{G}(V)$. Define Y^* as the set of ancestors of Y in $\mathcal{G}(V)$. Then $p(Y(f))$ is identified if and only if no element of A in D has children in D in $\mathcal{G}_D$. Moreover, if $p(Y(f))$ is identified, it is equal to*

$$\sum_{Y^* \setminus Y} \prod_{D \in \mathcal{D}(\mathcal{G}_{Y^*})} q_D(D | W_D)|_{\{A_i = f_i(A_i) : A_i \in A \cap \text{pa}(D)\}}$$

Proof: Fix disjoint subsets $A, Y \subseteq V$, define Y^* as the set of ancestors of Y in $\mathcal{G}(V)$, and assume for every $D \in \mathcal{D}(\mathcal{G}(V)_{Y^*})$, no element in $D \cap A$ has a child in D in $\mathcal{G}(V)_D$.

Fix a particular set of values of y^*, and for each $A_i \in A$, define $\tilde{a}_i \equiv f_i(y^*_{A_i})$, the value of a_i that $f_i \in f$ maps A_i to, if given the value of A_i in y^* as input. Define $\tilde{a} \equiv \{\tilde{a}_i : A_i \in A\}$. Note that if we can identify probabilities $p(Y^*(\tilde{a}) = y^*)$ for all values of y^*, we can obtain $p(Y(f))$ as a function of those probabilities.

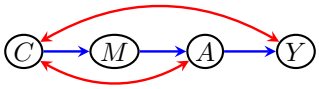
The fact that $p(Y^*(\tilde{a}) = y^*)$ is identified follows from Proposition 17 in Richardson and Robins (2013) applied to any hidden variable DAG $\mathcal{G}(V \cup H)$ yielding the latent projection $\mathcal{G}(V)$, as well as an inductive argument using Lemmas 52 and 55 that follows the proof of Theorem 60 in Richardson et al. (2017). The key observation here is that the proof of Theorem 60 (that establishes the soundness of the ID algorithm) never requires that sets A and Y remain disjoint to obtain identification of $p(Y(a))$, provided that the above precondition holds, specifically that no district containing $A_i \in A$ also contains a child of A.

To prove the converse, assume that some element in $D \cap A$ has a child in D in $\mathcal{G}(V)_D$. This immediately implies $p(D(a_i))$ is not identified, by the standard hedge construction. This then implies $p(D(a_i), A_i)$ is not identified, since otherwise $p(D(a_i))$ would be. This further implies that $p(Y(a_i), A_i)$ is not identified, by a standard argument based on one to one mappings on a subgraph $\mathcal{G}^*(V)$ containing D, a subset Y' of Y, and a set of directed paths from D to Y'. See, for instance, the proof of Theorem 6 in Shpitser and Sherman (2018).

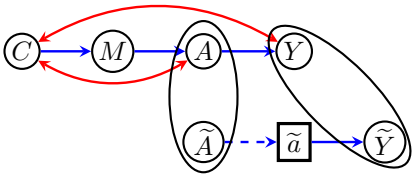
To show that $p(Y(A_i = f_i(A_i)))$ is not identified in $\mathcal{G}^*(V)$, we proceed as follows. Fix a directed path $A_i \to W_1 \to \ldots \to W_k \to Y_j \in Y'$, where $W_1 \in D$. Augment the graph $\mathcal{G}^*(V)$ with an extra set of copy vertices $\tilde{A}_i, \tilde{W}_1, \ldots, \tilde{W}_k, \tilde{Y}_j$, along with a edges forming a directed path from $\tilde{A}_i$ to $\tilde{Y}_j$ along these copy vertices, to yield a graph $\mathcal{G}^\dagger(V)$.

We now augment the two counterexample models showing non-identifiability of $p(Y(a_i), A_i)$ in $\mathcal{G}^*(V)$ to show $p(Y(a_i, \tilde{A}_i = \tilde{f}_i(\tilde{A}_i)), \tilde{Y}_j(a_i, \tilde{A}_i = \tilde{f}_i(\tilde{A}_i)))$ is not identified in $\mathcal{G}^\dagger(V)$. We do so by simply choosing $\tilde{f}_i$, as well as $p(\tilde{W}_1|\tilde{A}_i)$, $p(\tilde{W}_{i+1}|\tilde{W}_i)$ for $i = 1, \ldots, k-1$, and $p(\tilde{Y}_j|\tilde{W}_k)$ in both counterexample models to yield one to one mappings, as in the proof of Theorem 6 in Shpitser and Sherman (2018).

We conclude the proof by noting that by Lemma 1 in Shpitser and Sherman (2018), if $p(Y(a_i, \tilde{A}_i = \tilde{f}_i(\tilde{A}_i)), \tilde{Y}_j(a_i, \tilde{A}_i = \tilde{f}_i(\tilde{A}_i)))$ is not identified in any causal model represented by $\mathcal{G}^\dagger(V)$, then $p(Y(A_i = f_i^*(A_i)))$ is not identified in $\mathcal{G}^*(V)$, where variables $A_i \times \tilde{A}_i$, $W_i \times \tilde{W}_i$, for $i = 1, \ldots, k$, and $Y_j \times \tilde{Y}_j$ are treated as single variables, and a function f_i^* is defined as mapping $a_i' \times \tilde{a}_i$ values of $A_i \times \tilde{A}_i$ to $a_i \times \tilde{f}_i(\tilde{a}_i)$. □



We illustrate the proof of non-identifiability with the following example, where we are interested in identifying the SIP $p(Y(A = f(A)))$. In the above graph, A has a child Y in the district $\{A, C, Y\}$, hence our SIP should not be identified. To show this, note that $p(Y(a), C(a))$ is not identified, which is witnessed by a hedge containing sets $\{C, Y\}$, and $\{C, A, Y\}$. We now augment the graph $\mathcal{G}(V) = \mathcal{G}^*(V)$ above with extra vertex copies $\tilde{A}$, and $\tilde{Y}$, and an appropriate function $\tilde{f}_{\tilde{A}}$ represented below by a dashed edge from $\tilde{A}$ to $\tilde{a}$. We extend the proof of non-identification from the graph above to the graph below, and finally treat vertices $A, \tilde{A}$ and $Y, \tilde{Y}$ as a single vertex by taking a Cartesian product.



Theorem 4 *Fix disjoint subsets $A, Y \subseteq V$, and a set of* unrestricted *functions $f \equiv \{f_i : \mathfrak{X}_{A_i} \mapsto \mathfrak{X}_{A_i} | A_i \in A\}$ in a functional causal model given by the DAG $\mathcal{G}(V \cup H)$ that yields the latent projection ADMG $\mathcal{G}(V)$. Fix the set of all directed paths π in $\mathcal{G}(V)$ which start with $A_i \in A$, end in some element in $A \cup Y$, and which do not intersect elements in $A \cup Y$ otherwise. Define Y^* as the set of ancestors of Y in $\mathcal{G}(V)$, and $\tilde{Y}$ is the set of variables not in A which lie on a path in π that ends in Y. Then $p(Y(f(A)))$ is identified if and only if*

- *There are no two paths in π which start with the same edge, and where one path ends in an element of Y, and another in an element of A.*
- *Every element of A that lies in a district D in $\mathcal{G}(V)_{Y^*}$ does not have children in D in $\mathcal{G}_D$.*
- *For any two paths in π where the second vertex on the path is in D, either both paths have the final element in A or both paths have the final element in Y.*

Moreover, if $p(Y(f(A)))$ is identified, it is equal to

$$\sum_{Y^* \setminus Y} \prod_{D \in \mathcal{D}(\mathcal{G}_{Y^*})} q_D(D|W_D)|_{\{A_i = f_i(A_i) : A_i \in A \cap \mathrm{pa}^Y(D)\}},$$

where $\mathrm{pa}^Y(D)$ are parents of D along edges that are first edges on paths in π that end in Y.

Proof: Assume the preconditions above that allow identification hold. To simplify the identification argument, we consider extended graphs, as described in Malinsky et al. (2019). We construct an extended graph $\mathcal{G}^e$ for $\mathcal{G}(V)$, as follows. $\mathcal{G}^e$ contains all vertices in V. In addition, for each $A_i \in A$ that is ancestral of both A and Y, we construct two copies A_i^A, and A_i^Y that inherit the children of A as follows. A_i^A inherits children of A_i along edges that start directed paths from A_i into A, while A_i^Y inherits children of A_i along edges that start directed paths from A_i into Y. Vertices A_i^A and A_i^Y only have A_i as a parent. A_i itself only has A_i^A and A_i^Y as children. The structural equations corresponding to $\mathcal{G}^e$ are inherited from the models represented by $\mathcal{G}(V)$, except the structural equations for added variables A_i^A and A_i^Y, which are identity functions.

The advantage of extended graphs is they allow us to rephrase mediation analysis problems defined in terms

of edge interventions in terms of standard intervention operations on added copy variables. This was shown in Propositions 9 and 10 in Malinsky et al. (2019). In our case, we proceed as follows.

Let $A^Y \equiv \{A_i^Y : A_i \in A\}$. Fix a particular set of values of y^* in Y^* in $\mathcal{G}^e$, and for each $A_i^Y \in A^Y$, define $\tilde{a}_i \equiv f_i(y^*_{A_i^Y})$, the value of a_i that $f_i \in f$ maps A_i^Y to, if given the value of A_i^Y in y^* as input. Define $\tilde{a}^Y \equiv \{\tilde{a}_i^Y : A_i^Y \in A^Y\}$. Note that if we can identify probabilities $p(Y^*(A^Y = \tilde{a}^Y) = y^*)$ for all values of y^* in $\mathcal{G}^e$, we can obtain $p(Y(f(A)))$ as a function of those probabilities. Note that the counterfactual $Y^*(A^Y = \tilde{a}^Y)$ only intervenes on elements of A^Y, and leaves all A_i^A for $A_i \in A$ at their natural values. The fact that $p(Y^*(A^Y = \tilde{a}^Y) = y^*)$ is identified, given the preconditions, follows from Theorem 4, and the last pre-condition, which guarantees each intrinsic kernel $q_D(D|W_D)$ has positive support in the observed data distribution.

Assume the preconditions above that allow identification do not hold. If the first precondition does not hold, non-identification follows by Theorem 2, even in the edge subgraph of $\mathcal{G}(V)$ containing only directed edges, and thus also in $\mathcal{G}(V)$.

If the first precondition holds, but the second precondition does not hold, fix a district D containing A_i and a child of A_i in $\mathcal{G}_D$. We then consider a SIT that sets every element in $A \setminus \{A_i\}$ to their natural value (which is allowed since the set f is unrestricted functions), and follow the proof of non-identification in Theorem 3.

If the first two preconditions hold, but the last precondition does not hold, fix A_i with two edges into a district D with the required property. Then consider the SIT that sets every element in $A \setminus \{A_i\}$ to their natural value, and that sets A_i to a_i. The distribution $p(Y(A_i^Y = a_i))$ is not identified in $\mathcal{G}^e$ as a simple corollary of Theorem 5 in Shpitser and Sherman (2018), and Proposition 10 in Malinsky et al. (2019). To see that $p(Y(A_i^Y = f_i(A_i^Y)))$ is also not identified in $\mathcal{G}^e$, we follow the argument in Theorem 3 that uses vertex copies and Cartesian products. □

Theorem 5 *Fix* $\beta = \sum_{C,A} \mathbb{E}[Y|a = f(A), C]p(A|C)p(C)$*, which is equal to* $\mathbb{E}[Y(f(A))] = \mathbb{E}[Y(f)]$ *under the model in Fig. 1 (a). The efficient influence function for* β *under the non-parametric observed data model is given by*

$$U(\beta) = \frac{\sum_{A'} \mathbb{I}(A = f(A'))p(A' \mid C)}{p(A \mid C)}\{Y - \mathbb{E}[Y \mid A, C]\} + \mathbb{E}[Y \mid a = f(A), C)] - \beta \tag{6}$$

Proof: The model imposes no restrictions on the observed data distribution, hence it is non-parametric saturated and has a unique (and thus efficient) influence function for β. This influence function is given by the solution to the following integral equation

$$\frac{\partial}{\partial t}\beta(F_t)|_{t=0} = E[S(C, A, Y)\psi(\beta)],$$

where $S(C, A, Y)$ is the observed data score.

Using the product rule of differentiation, we have

$$\begin{aligned}\frac{\partial}{\partial t}\beta(F_t) = &\sum_{C,A,Y} y\frac{\partial}{\partial t}p(Y \mid a = f(A), C)p(A \mid C)p(C) \\ &+ \sum_{C,A,Y} yp(Y \mid a = f(A), C)\frac{\partial}{\partial t}p(A \mid C)p(C) \\ &+ \sum_{C,A,Y} yp(Y \mid a = f(A), C)p(A \mid C)\frac{\partial}{\partial t}p(C)\end{aligned}$$

Starting with the first term and using properties of scores and the chain rule of differentiation, we have

$$\sum_{C,A,Y} yS(Y|a = f(A), C)p(Y|a = f(A), C)p(A|C)p(C)$$

We introduce an indicator and multiply and divide by $p(A'|C)$ to obtain

$$= \sum_A \mathbb{E}_{Y,A',C}\left[\frac{\mathbb{I}(A' = f(A))}{p(A'|C)}p(A|C)yS(Y|A', C)\right]$$

Then, introducing a term that is constant w.r.t. the conditional score $S(Y \mid A', C)$, and using the tower law gives

$$\begin{aligned}&= \sum_A \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{I}(A' = f(A))}{p(A' \mid C)}p(A|C)\{y - \mathbb{E}[Y|A', C]\}S(Y|A', C)|A', C\right]\right] \\ &= \sum_A \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{I}(A' = f(A))}{p(A' \mid C)}p(A|C)\{y - \mathbb{E}[Y|A', C]\}S(Y, A', C)\right]\right] \\ &= \mathbb{E}\left[\sum_A \frac{\mathbb{I}(A' = f(A))}{p(A'|C)}p(A|C)\{y - \mathbb{E}[Y|A', C]\}S(Y, A', C)\right]\end{aligned}$$

The contribution to the influence function from the first term is therefore

$$U_1(\beta) = \sum_A \frac{\mathbb{I}(A' = f(A))}{p(A' \mid C)}p(A \mid C)\{y - \mathbb{E}[Y \mid A', C]\}$$

For the second term, we have

$$\begin{aligned}&\sum_{C,A,Y} yp(Y \mid a = f(A), C)S(A \mid C)p(A \mid C)p(C) \\ &= \sum_{C,A} \mathbb{E}[Y \mid a = f(A), C]S(A \mid C)p(A \mid C)p(C)\end{aligned}$$

Introducing a term constant w.r.t. $S(A \mid C)$ gives

$$\sum_{C,A}\left(\mathbb{E}[Y|a=f(A),C]-\sum_{A}\mathbb{E}[Y|a=f(A),C)]p(A|C)\right)$$
$$S(A|C)p(A|C)p(C).$$

Since the term in $(\cdot)$ is mean zero given C, we can introduce the required score $S(C)$ to obtain

$$\sum_{C,A}\left(\mathbb{E}[Y \mid a=f(A),C]-\sum_{A}\mathbb{E}[Y \mid a=f(A),C]p(A \mid C)\right)$$
$$S(A,C)p(A,C)$$

Since the above expression is not a function of Y, so we can sum over Y to obtain

$$\sum_{C,A,Y}\left(\mathbb{E}[Y|a=f(A),C]-\sum_{A}\mathbb{E}[Y|a=f(A),C]p(A|C)\right)$$
$$S(Y,A,C)p(Y,A,C)$$

The contribution to the influence function from the second term is therefore

$$U_2(\beta)=\mathbb{E}[Y|a=f(A),C)]-\sum_{A}\mathbb{E}[Y|a=f(A),C]p(A|C)$$

Moving on to the third term, we have

$$\begin{aligned}
&\sum_{C,A,Y} Yp(Y|a=f(A),C)p(A|C)\frac{\partial p(C)}{\partial dt}\\
&=\sum_{C,A,Y} Yp(Y|a=f(A),C)p(A|C)S(C)p(C)\\
&=\sum_{C,A}\mathbb{E}[Y|a=f(A),C]p(A|C)S(C)p(C)\\
&=\mathbb{E}_C\left[\sum_{A}E[Y|a=f(A),C]p(A|C)S(C)\right]
\end{aligned}$$

We can now introduce a term independent of $S(C)$:

$$\begin{aligned}
&\mathbb{E}_C[\{\sum_{A}\mathbb{E}[Y|a'=f(a),C]p(A|C)\\
&-\mathbb{E}_C[\sum_{A}E[Y|a=f(A),C]p(A|C)]\}S(C)]\\
&=\mathbb{E}_C[\mathbb{E}_{Y,A|C}[\{\sum_{A}\mathbb{E}[Y|a=f(A),C]p(A|C)\\
&-\mathbb{E}_C[\sum_{A}E[Y|a=f(A),C]p(A|C)]\}S(C)|C]]\\
&=\mathbb{E}_C[E_{Y,A|C}[\{\sum_{A}\mathbb{E}[Y|a=f(A),C]p(A|C)\\
&-\mathbb{E}_C[\sum_{A}E[Y|a=f(A),C]p(A|C)]\}\{S(C)\\
&+S(Y,A|C)\}|C]]\\
&=\mathbb{E}_{Y,A,C}[\{\sum_{A}\mathbb{E}[Y|a=f(A),C]p(A|C)\\
&-\mathbb{E}_C[\sum_{A}E[Y|a=f(A),C]p(A|C)]\}S(Y,A,C)],
\end{aligned}$$

which implies the contribution to the influence function of the third term is

$$U_3(\beta)=\sum_{A}E[Y|a=f(A),C]p(A|C)-\beta$$

Putting all three terms together, the final influence function is:

$$\begin{aligned}
U(\beta)=&\sum_{A}\frac{\mathbb{I}(A'=f(A))}{p(A' \mid C)}p(A \mid C)\{Y-\mathbb{E}[Y \mid A',C]\}\\
&+\mathbb{E}[Y \mid a=f(A),C)]-\beta
\end{aligned}$$

□

Theorem 6 *The estimator of β which solves the estimating equation $\mathbb{E}[U(\beta)]=0$ is consistent, asymptotically normal (CAN) in the union model where one of $\pi(C;\eta_A)=p(A|C)$, $m(A,C;\eta_Y)=\mathbb{E}[Y|A,C]$ is correctly specified.*

Proof: Assume $p(A \mid C)$ is specified incorrectly as $p^*(A \mid C)$. Then the estimator for β is given as

$$\begin{aligned}
\mathbb{E}[U(\beta)]=&\sum_{A}\frac{\mathbb{I}(A'=f(A))}{p^*(A' \mid C)}p^*(A \mid C)\{Y-\mathbb{E}[Y \mid A',C]\}\\
&+\mathbb{E}[Y \mid a=f(A),C)]-\beta]
\end{aligned}$$

Since $U(\beta)$ is linear in β, we can solve for β explicitly as an expectation with two terms. The first term is

$$\mathbb{E}\left[\sum_{A}\frac{\mathbb{I}(A'=f(A))}{p^*(A' \mid C)}p^*(A \mid C)\{Y-\mathbb{E}[Y \mid A',C]\}\right]$$

which is mean zero, since $\mathbb{E}[Y|A,C]$ is specified correctly. The second term is $\mathbb{E}[\mathbb{E}[Y \mid a=f(A),C)]]$ which is equal to β by definition if $\mathbb{E}[Y|A,C]$ is specified correctly.

Assume $\mathbb{E}[Y \mid A,C]$ is specified incorrectly as $\mathbb{E}^*[Y \mid A,C]$. The estimator for β is then

$$\begin{aligned}
\mathbb{E}[U(\beta)]=&\frac{\sum_A \mathbb{I}(A'=f(A))p(A \mid C)}{p(A' \mid C)}\{Y-\mathbb{E}^*[Y \mid A',C]\}\\
&+\mathbb{E}^*[Y \mid a=f(A),C)]-\beta]
\end{aligned}$$

which can be rewritten as a sum of two terms. The first is

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}^*[Y|a=f(A),C)]-\\
&\frac{\sum_A \mathbb{I}(A'=f(A))p(A|C)}{p(A'|C)}\mathbb{E}^*[Y|A',C]],
\end{aligned}$$

which is mean zero. The second term is

$$\mathbb{E}\left[\frac{\sum_A \mathbb{I}(A'=f(A))p(A|C)}{p(A'|C)}Y\right],$$

which evaluates to β if $p(A|C)$ is correctly specified.

All of the above estimators are special cases of the RAL estimator for β based on the efficient influence function. As a result, standard regularity assumptions Robins et al. (1992), and properties of maximum likelihood estimators imply both estimators are CAN. □

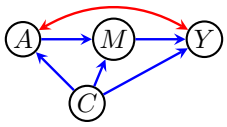


Figure 4: The front-door model with a vector of baseline covariates C

Theorem 7 *Fix* $\beta = \sum_{C,M,A} \mathbb{E}[Y|a=A,C]p(A,C)p(M \mid a=f(A),C)$*, which is equal to* $\mathbb{E}[Y(f(A))] = \mathbb{E}[Y(f)]$ *under the model in the figure 4. The efficient influence function for* β *under the non-parametric observed data model is given by*

$$U(\beta) = \frac{p(M \mid a=f(A),C)}{p(M \mid A,C)}\{Y - \mathbb{E}[Y \mid M,A,C]\} \quad (7)$$
$$+ \sum_{M} \mathbb{E}[Y \mid M,A,C]p(M \mid a=f(A),C) - \beta$$
$$\sum_{A'} \frac{I(A=f(A'))}{p(A \mid C)}\{q - \sum_{M} q \cdot p(M \mid A,C)\}$$

where $q \equiv \mathbb{E}[Y \mid M,A',C]p(A' \mid C)$.

Proof: The model imposes no restrictions on the observed data distribution, hence it is non-parametric saturated and has a unique (and thus efficient) influence function for β. This influence function is given by the solution to the following integral equation

$$\frac{\partial}{\partial t}\beta(F_t)|_{t=0} = E[S(C,A,M,Y)\psi(\beta)],$$

where $S(C,A,Y)$ is the observed data score.

Using the product rule of differentiation, we have

$$\frac{\partial}{\partial t}\beta(F_t) = \sum_{Y,C,M,A} Y \frac{\partial p(Y \mid M,A,C)}{\partial t}p(A,C)p(M \mid a=f(A),C)$$
$$\sum_{C,M,A} \mathbb{E}[Y \mid M,A,C]\frac{\partial p(A,C)}{\partial t}p(M \mid a=f(A),C)$$
$$\sum_{C,M,A} \mathbb{E}[Y \mid M,A,C]p(A,C)\frac{\partial p(M \mid a=f(A),C)}{\partial t}$$

Starting with the first term and using properties of scores and the chain rule of differentiation, we have

$$\sum_{Y,C,M,A} YS(Y \mid M,A,C)p(Y \mid M,A,C)$$
$$p(A,C)p(M \mid a=f(A),C)$$

Multiplying and diving by $p(M \mid A,C)$, we get

$$\sum_{C,M,A} \frac{p(M \mid a=f(A),C)}{p(M \mid A,C)}\mathbb{E}[Y \mid M,A,C]S(Y \mid M,A,C)p(M,A,C)$$
$$= \mathbb{E}_{Y,M,A,C}[\frac{p(M \mid a=f(A),C)}{p(M \mid A,C)}YS(Y \mid M,A,C)]$$

Since

$$\mathbb{E}[\frac{p(M \mid a=f(A),C)}{p(M \mid A,C)}\mathbb{E}[Y \mid M,A,C]S(Y \mid M,A,C)] = 0,$$

the previous equation can be rewritten as

$$\mathbb{E}[\frac{p(M \mid a=f(A),C)}{p(M \mid A,C)}\{Y - \mathbb{E}[Y \mid M,A,C]\}S(Y \mid M,A,C)]$$

Finally, using

$$\mathbb{E}[\frac{p(M \mid a=f(A),C)}{p(M \mid A,C)}\{Y - \mathbb{E}[Y \mid M,A,C]\}S(M,A,C)] = 0$$

gives us

$$\mathbb{E}[\frac{p(M \mid a=f(A),C)}{p(M \mid A,C)}\{Y - \mathbb{E}[Y \mid M,A,C]\}S(Y,M,A,C)]$$

The contribution to the influence function from the first term is

$$U_1(\beta) = \frac{p(M \mid a=f(A),C)}{p(M \mid A,C)}\{Y - \mathbb{E}[Y \mid M,A,C]\}$$

For the second term, we have:

$$\sum_{C,M,A} \mathbb{E}[Y \mid M,A,C]S(A,C)p(A,C)p(M \mid a=f(A),C)$$
$$= \mathbb{E}_{A,C}[\sum_{M} \mathbb{E}[Y \mid M,A,C]p(M \mid a=f(A),C)S(A,C)]$$

Using $[\mathbb{E}_{A,C}[\sum_M \mathbb{E}[Y \mid M,A,C]p(M \mid a=f(A),C)]S(A,C)] = 0$, along with properties of the score, the above is equal to

$$\mathbb{E}[\{\sum_{M} \mathbb{E}[Y \mid M,A,C]p(M \mid a=f(A),C) - \beta\}S(Y,M,A,C)]$$

The contribution of the second term to the influence function is:

$$U_2(\beta) = \sum_{M} \mathbb{E}[Y \mid M,A,C]p(M \mid a=f(A),C) - \beta$$

Finally, moving on to the third term. Using similar steps as before, the third term can be written as

$$\sum_{C,M,A} \mathbb{E}[Y \mid M,A,C]p(A,C)$$
$$S(M \mid a=f(A),C)p(M \mid a=f(A),C)$$

Introducing another random variable A' distributed as A, and using the indicator function

$$\sum_{C,M,A,A'} I(A' = f(A))\mathbb{E}[Y \mid M, A, C]p(A, C)S(M \mid A', C)p(M \mid A', C)$$

$$\sum_{C,M,A,A'} \frac{I(A' = f(A))}{p(A' \mid C)}\mathbb{E}[Y \mid M, A, C]p(A \mid C)S(M \mid A', C)p(M, A', C)$$

$$\mathbb{E}_{C,M,A'}[\sum_{A} \frac{I(A' = f(A))}{p(A' \mid C)}\mathbb{E}[Y \mid M, A, C]p(A \mid C)S(M \mid A', C)]$$

Interchanging the labels of A' and A for clarity, the previous expectation is rewritten as:

$$\mathbb{E}_{C,M,A}[\sum_{A'} \frac{I(A = f(A'))}{p(A \mid C)}\mathbb{E}[Y \mid M, A', C]p(A' \mid C)S(M \mid A, C)]$$

Denoting $\mathbb{E}[Y \mid M, A', C]p(A' \mid C)$ by q, and utilizing $\mathbb{E}[\sum_{A'} \frac{I(A=f(A'))}{p(A|C)} \sum_M qp(M \mid A, C)S(M \mid A, C)] = 0$, we have

$$\begin{aligned}
&\mathbb{E}[\sum_{A'} \frac{I(A = f(A'))}{p(A \mid C)}\{q - \sum_{M} qp(M \mid A, C)\}S(M \mid A, C)] \\
=&\mathbb{E}[\sum_{A'} \frac{I(A = f(A'))}{p(A \mid C)}\{q - \sum_{M} qp(M \mid A, C)\}S(M, A, C)] \\
=&\mathbb{E}[\mathbb{E}[\sum_{A'} \frac{I(A = f(A'))}{p(A \mid C)}\{q - \sum_{M} qp(M \mid A, C)\}\{S(Y, M, A, C)\} \mid M, A, C]] \\
=&\mathbb{E}[\sum_{A'} \frac{I(A = f(A'))}{p(A \mid C)}\{q - \sum_{M} qp(M \mid A, C)\}S(Y, M, A, C)]
\end{aligned}$$

Hence the contribution of the third term to the influence function is:

$$U_3(\beta) = \sum_{A'} \frac{I(A = f(A'))}{p(A \mid C)}\{q - \sum_{M} qp(M \mid A, C)\}$$

Putting it all together, the influence function is given by:

$$\begin{aligned}
U(\beta) =& \frac{p(M \mid a = f(A), C)}{p(M \mid A, C)}\{Y - \mathbb{E}[Y \mid M, A, C]\} \\
&+ \sum_{M} \mathbb{E}[Y \mid M, A, C]p(M \mid a = f(A), C) - \beta \\
&\sum_{A'} \frac{I(A = f(A'))}{p(A \mid C)}\{q - \sum_{M} qp(M \mid A, C)\}
\end{aligned}$$

□