16

Identification In Graphical Causal Models

Ilya Shpitser

Department of Computer Science, Johns Hopkins University

CONTENTS

16.1 Introduction

→ Undirected: global / pairwise Markov property

Previous chapters introduced statistical graphical models: sets of distributions defined by conditional independences linked, via Markov properties, to absences of edges in directed and mixed graphs.

→ Directed: global / local Markov property

Directed acyclic graphs (DAGs) have also been extended to represent causal models—a viewpoint developed in the preceding Chapter 15. Like statistical models, causal models can be viewed as sets of distributions defined by certain restrictions. However, unlike statistical models, a causal model also includes distributions that are counterfactual, rather than factually observed. Random variables following such counterfactual distributions, called potential outcomes, correspond to results of hypothetical experiments by means of which causality is represented. Restrictions given by causal models provide a link between factual and counterfactual distributions. This link allows identification of certain distributions of potential outcomes from the observed data distribution, and ultimately allows causal inferences to be made from data.

→ similarity
→ difference
potential outcome ←
↳ presence / absence of edges

identification theory

In this chapter, we will describe a simple formulation of identification theory for common targets of inference that arise in causal inference, developed in the context of non-parametric graphical causal models. In addition, we will describe extensions of this theory to an important type of causal model where counterfactual random variables are determined via linear causal mechanisms and Gaussian noise. These models are known as linear structural equation models with correlated errors.

→ general non-parametric

↳ parametric: linear causal mechanism + Gaussian noise

16.2 Causal Models Of A DAG

Throughout this section, fix an arbitrary DAG $\mathcal{G}$ with a vertex set $\mathbf{V}$ whose elements have a one to one correspondence to the random variables under consideration. When necessary, we will denote $\mathcal{G}$ with an explicit vertex set as $\mathcal{G}(\mathbf{V})$. The statistical model of $\mathcal{G}$, or a Bayesian network, is the set of all joint distributions $p(\mathbf{V})$ obeying the following factorization restriction

$$p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V \mid \mathrm{pa}_{\mathcal{G}}(V)), \tag{16.1}$$

→ existence of edges

where $\mathrm{pa}_{\mathcal{G}}(V)$, also known as the set of *parents* of V, is defined as $\{W \in \mathbf{V} \mid W \to V \text{ exists in } \mathcal{G}\}$. In cases where $\mathcal{G}$ is clear, we will omit $\mathcal{G}$ to yield $\mathrm{pa}(V)$. Such distributions are also said to be Markov relative to $\mathcal{G}$.

Factorization → Markov property

Causal models of a DAG $\mathcal{G}$ are also defined via restrictions on sets of joint distributions. Unlike statistical models, however, causal models consist of joint distributions over *potential outcome* random variables, defined via an intervention operation. Let $\mathbf{A} \subset \mathbf{V}$ be a set of random variables, and let $\mathbf{a}$ be a set of possible values for the variables in $\mathbf{A}$. Then $\mathrm{do}(\mathbf{a})$ denotes an intervention in which an experimenter controls the variables in $\mathbf{A}$ and sets them to the values determined by $\mathbf{a}$. If $Y \in \mathbf{V}$ is another variable then $Y(\mathbf{a})$ is used to denote the response of Y to the intervention $\mathrm{do}(\mathbf{a})$. These types of random variables are called potential outcomes, because Y is often an outcome of interest, and the intervention is often hypothetical, rather than actually occurring. To represent causality on a DAG $\mathcal{G}$, we will encode the outcomes of every possible intervention operation, starting with interventions on sets $\mathrm{pa}(V)$ for every $V \in \mathbf{V}$. For all value assignments $\mathbf{a}$ to $\mathrm{pa}(V)$, we assume the existence of the potential outcome $V(\mathbf{a})$, and a well-defined joint distribution over these random variables. One way to justify the existence of $V(\mathbf{a})$ is by postulating a set of *structural equations* $\{f_V(\mathbf{a}, \epsilon_V) \mid V \in \mathbf{V}\}$. Each function f_V maps a set of values $\mathbf{a}$ of parents of V in $\mathcal{G}$, and a random variable ϵ_V representing exogenous factors not captured by the model, to values of V. These functions act as causal mechanisms, and are assumed to be invariant to intervention operations. For any value set $\mathbf{a}$, the function f_V, and the exogenous variable ϵ_V induce the random variable $V(\mathbf{a})$.

We use these potential outcomes, and the associated joint distributions, to define other potential outcomes using *recursive substitution*. For any $\mathbf{A} \subseteq \mathbf{V}$, and any values $\mathbf{a}$ of $\mathbf{A}$, we define for every $Y \in \mathbf{V}$

$$Y(\mathbf{a}) \equiv Y(\mathbf{a}_{\mathrm{pa}(Y) \cap \mathbf{A}}, \{W(\mathbf{a}) \mid W \in \mathrm{pa}(Y) \setminus \mathbf{A}\}) \tag{16.2}$$

In words, this states that the response of Y to $\mathrm{do}(\mathbf{a})$ is defined as the potential outcome where all parents of Y which are in $\mathbf{A}$ are assigned an appropriate value from $\mathbf{a}$, and all other parents $W \in \mathrm{pa}(Y) \setminus \mathbf{A}$ are assigned whatever value they would have attained under $\mathrm{do}(\mathbf{a})$. These are defined recursively, and the definition terminates because of the lack of

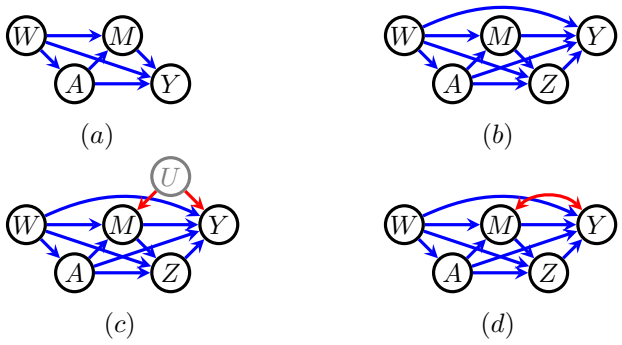


FIGURE 16.1: (a) A causal DAG with a single treatment A, a single outcome Y, a vector W of baseline variables, and a single mediator M. (b) A more complex causal DAG with two mediators M and Z. (c) A version of the DAG in (b) with an unobserved confounder U. (d) The ADMG obtained from the DAG in (c) via the latent projection operation collapsing over the unobserved variable U.

↳ acyclic directed mixed graph

directed cycles in $\mathcal{G}$. For example, in the graph in Fig. 16.1 (a), $Y(a) = Y(a, M(a, W), W)$. The distribution $p(\{Y(\mathbf{a}) \mid Y \in \mathbf{Y}\})$ is sometimes written as $p(\mathbf{Y} \mid \text{do}(\mathbf{a}))$, as in [16].

Just as a statistical model of $\mathcal{G}$ is a set of distributions over $\mathbf{V}$ defined by (16.1), a causal model of $\mathcal{G}$ is a set of distributions over random variables in the set

$$\{V(\mathbf{v}_{\text{pa}(V)}) \mid V \in \mathbf{V}, \text{ any set of values } \mathbf{v} \text{ of } \mathbf{V}\}$$

↳ $p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V \mid \text{pa}_{\mathcal{G}}(V))$,

defined by some restrictions. We consider two such models, described in [21]. The *finest fully randomized causally interpretable structured tree graph model (FFRCISTGM)*, or the *single world model* [29], is the set of all distributions such that variables in

↳ restriction ①

$$\{V(\mathbf{v}_{\text{pa}(V)}) \mid V \in \mathbf{V}\}$$

are mutually independent for every set of values $\mathbf{v}$ of $\mathbf{V}$. The *non-parametric structural equation model with independent errors (NPSEM-IE)*, also known as the *functional model*, or the *multiple worlds model* [29], is the set of all distributions such that variables in

restriction ②

$$\{\{V(\mathbf{a}_{\text{pa}(V)}) \mid \mathbf{a}_{\text{pa}(V)} \text{ any set of values of pa}(V)\} | V \in \mathbf{V}\} \tag{16.4}$$

are mutually independent. The multiple worlds model associated with $\mathcal{G}$ is a submodel of the single world model associated with $\mathcal{G}$, because it always places at least as many restrictions on potential outcome responses, and in most cases many more. Specifically, the single world model only imposes independence restrictions within single hypothetical worlds, as specified by consistent interventions do($\mathbf{v}$), for any set of values $\mathbf{v}$ of $\mathbf{V}$. By contrast, the multiple worlds model, in addition, imposes independence restrictions across multiple hypothetical worlds.

single-hypothesis

For example, the binary single world model associated with the DAG in Fig. 16.1 (a) asserts that variables W, $A(w)$, $M(a, w)$, $Y(a, m, w)$ are mutually independent for any $a, m, w \in \{0, 1\}$, while the binary multiple worlds model associated with the same DAG asserts that sets $\{W\}, \{A(w) \mid w \in \{0, 1\}\}, \{M(a, w) \mid a \in \{0, 1\}, w \in \{0, 1\}\}, \{Y(a, m, w) \mid a \in \{0, 1\}, m \in \{0, 1\}, w \in \{0, 1\}\}$ are mutually independent. As their names imply, the single world model only imposes restrictions on a set of variables under a single consistent set of interventions, while the multiple worlds model may also impose restrictions on variables

→ multiple-hypothesis

across multiple conflicting sets of interventions simultaneously. If the set of random variables in (16.3) is viewed as defined by a set of structural equations

$$\{f_V(\mathbf{v}_{\mathrm{pa}(V)}, \epsilon_V) \mid V \in \mathbf{V}, \text{ any set of values } \mathbf{v} \text{ of } \mathbf{V}\},$$

then the assumption (16.4) corresponding to the multiple worlds model can be interpreted to mean that the distribution $p(\{\epsilon_V \mid V \in \mathbf{V}\})$ factorizes as $\prod_{V \in \mathbf{V}} p(\epsilon_V)$.

16.2.1 Causal, Direct, Indirect, and Path-Specific Effects

Targets of inference in causal inference are functions of potential outcomes. Chapter 15 has already described counterfactual mean contrasts such as the average causal effect (ACE):

$$\mathbb{E}[Y(\mathbf{a})] - \mathbb{E}[Y(\mathbf{a}')].$$

which quantify the overall causal effect of a set of treatments $\mathbf{A}$ on the outcome Y, and the direct and indirect effects on the difference scale [23, 15]:

$$\begin{aligned}\mathbb{E}[Y(a)] - \mathbb{E}[Y(a')] &= \underbrace{(\mathbb{E}[Y(a)] - \mathbb{E}[Y(a, M(a'))])}_{\text{total indirect effect}} + \underbrace{(\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a')])}_{\text{pure direct effect}} \\ &= \underbrace{(\mathbb{E}[Y(a)] - \mathbb{E}[Y(a', M(a))])}_{\text{total direct effect}} + \underbrace{(\mathbb{E}[Y(a', M(a))] - \mathbb{E}[Y(a')])}_{\text{pure indirect effect}}\end{aligned}$$

which quantify the extent to which the causal effect of A on Y is mediated by a third variable M.

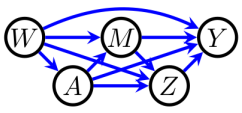
Direct and indirect effects are defined using nested potential outcomes such as $Y(a, M(a'))$. These variables represent hypothetical situations where A is set to one value for the purposes of the direct causal path $A \to Y$ from treatment to outcome, and to another value for the purposes of the indirect causal path $A \to M \to Y$ mediated by M. These different treatment settings are crucial for defining effects which isolate the causal influence along one but not the other path.

The intuition behind direct and indirect effects can be generalized to settings where an effect along a particular causal path that is neither direct nor through a single mediator M is of interest. Effects of exposure on outcome along a predefined set of paths are called path-specific effects [15]. To define such effects we need to define a potential outcome Y where the treatment is set to one value a with respect to the set of paths of interest, and to another value a' with respect to all other paths. This is a natural generalization of the way direct effects were defined above, with the direct path $A \to Y$ being a path of interest, and the indirect path $A \to M \to Y$ being "all other paths." We can modify the recursive substitution definition (16.2) to define such potential outcomes as follows. Given a set of directed paths π from A to Y, and values a, a', define the π-specific potential outcome Y as

$$\begin{aligned}Y(\pi, a, a') &\equiv a \text{ if } Y = A \\ Y(\pi, a, a') &\equiv Y(\{W(\pi, a, a') \mid W \in \mathrm{pa}^{\pi}(Y)\}, \{W(a') \mid W \in \mathrm{pa}^{\bar{\pi}}(Y)\})\end{aligned} \tag{16.5}$$

where $W(a') \equiv a'$ if $W = A$, $\mathrm{pa}^{\pi}(Y)$ is the set of parents of Y along an edge which is a part of a path in π, and $\mathrm{pa}^{\bar{\pi}}(Y)$ is the set of all other parents of Y. Applying this definition to π consisting of a single path $A \to Z \to Y$ in Fig. 16.1 (b) yields $Y(a', Z(a, M(a', W)), M(a', W), W)$. By analogy with direct and indirect effects, we can use this counterfactual to define the total effect not through π as

$$\begin{aligned}&\mathbb{E}[Y(a)] - \mathbb{E}[Y(\{A \to Z \to Y\}, a, a')] = \\ &\mathbb{E}[Y(a)] - \mathbb{E}[Y(a', Z(a, M(a', W)), M(a', W), W)].\end{aligned}$$



(b)

$$\underbrace{(\mathbb{E}[Y(a)] - \mathbb{E}[Y(a, M(a'))])}_{\text{total indirect effect}} + \underbrace{(\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a')])}_{\text{pure direct effect}}$$

and the pure π-specific effect as

$$\mathbb{E}[Y(\{A \to Z \to Y\}, a, a')] - \mathbb{E}[Y(a')] = \\ \mathbb{E}[Y(a', Z(a, M(a', W)), M(a', W), W)] - \mathbb{E}[Y(a')].$$

Equation (16.5) generalizes in a natural way to a set of treatments $\mathbf{A}$, and multiple outcomes $\mathbf{Y}$. In such cases, attention is restricted to *proper causal paths* for $\mathbf{A}$ and $\mathbf{Y}$, which are directed paths from an element in $\mathbf{A}$ to an element in $\mathbf{Y}$ that otherwise do not intersect $\mathbf{A}$.

16.2.2 Responses To Dynamic Treatment Regimes

In settings such as precision medicine, the primary goal is obtaining good outcomes for every individual, rather than assessing average treatment effects in a population, of the kind defined above. In simple versions of this setting, shown in Fig. 16.1 (a), the goal is to use data generated from the observed data distribution

$$p(W)\, p(A|W)\, p(M|A, W)\, p(Y|M, A, W),$$

→ opposed to atomic-intervention

where the treatment A can be viewed as assigned based on a (likely suboptimal, possibly stochastic) policy represented by $p(A|W)$, to infer a "better" hypothetical policy $g_A(W)$. Here $g_A(W)$ is a mapping, taken from some class, from values of W to treatment values a. Such policies are sometimes known as *dynamic treatment regimes* [4]. Policy quality can be defined in a number of ways, but is often defined using expected outcomes given the policy: $\mathbb{E}[Y(A = g_A(W))]$, where the expectation is taken with respect to the distribution $p(W)$. Effective policies like g_A may assign a known primary treatment a for patients unless W suggests severe side effects for a, in which case an alternative a' is given.

More generally, given a set of treatments $\mathbf{A}$ in a causal model represented by a DAG $\mathcal{G}(\mathbf{V})$, fix a topological ordering $\prec$ on $\mathbf{V}$ consistent with $\mathcal{G}$, and consider for every $A \in \mathbf{A}$ a set of variables $\mathbf{W}_A$ earlier in the ordering $\prec$ than A. Fix a set of policies $\mathbf{g_A} \equiv \{g_A(\mathbf{W}_A) \mid A \in \mathbf{A}\}$ that determine the value of each $A \in \mathbf{A}$ using values of $\mathbf{W}_A$. Policies in $\mathbf{g_A}$ could be either deterministic or stochastic, depending on the application. For any $Y \in \mathbf{V} \setminus \mathbf{A}$, the counterfactual response Y had every $A \in \mathbf{A}$ been determined by $\mathbf{g_A}$ is defined using the natural generalization of the recursive substitution definition (16.2):

$$Y(\mathbf{g_A}) = Y(\{A = g_A(\mathbf{W}_A(\mathbf{g_A})) | A \in \mathrm{pa}(Y) \cap \mathbf{A}\}, \{W(\mathbf{g_A}) | W \in \mathrm{pa}(Y) \setminus \mathbf{A}\}).$$

$Y(\mathbf{a}) \equiv Y(\mathbf{a}_{\mathrm{pa}(Y)\cap\mathbf{A}}, \{W(\mathbf{a}) \mid W \in \mathrm{pa}(Y) \setminus \mathbf{A}\})$

As an example, in Fig. 16.1 (b), if we are interested in the potential outcome Y, had A and Z been set using policies $g_A(W)$ and $g_Z(M, W)$, this outcome would be

$$Y(\mathbf{g_A}) \equiv Y(A = g_A(W), Z = g_Z(M = M(\mathbf{g_A}), W), W) \\ M(\mathbf{g_A}) \equiv M(A = g_A(W), W). \tag{16.7}$$

These kinds of examples arises in management of cancer, HIV, or other chronic diseases where choices of primary therapy (such as induction chemotherapy for cancer patients) A depend on baseline patient characteristics such as age, and the choice to switch to a second line therapy Z (such as salvage chemotherapy) depends on intermediate outcomes M which themselves depend on the choice of primary therapy.

16.2.3 Identifiability

Causal models described so far are sets of joint distributions over random variables in the set $\{V(\mathbf{a}_{\mathrm{pa}(V)}) \mid V \in \mathbf{V}\}$, and other random variables defined from this set via (16.2), (16.5),

and (16.6). In particular, the *observed distribution* $p(\mathbf{V})$ is always a marginal of any joint distribution that is an element of a causal model. This distribution is important in causal inference applications since data drawn from this distribution is what is typically available.

A distribution $\widetilde{p}$ is said to be globally identified from $p(\mathbf{V})$ in a causal model, if there exists a function g such that $\widetilde{p} = g(p(\mathbf{V}))$ in every element of the model. Weaker notions of identifiability, such as generic identifiability where $\widetilde{p}$ is a function $g(p(\mathbf{V}))$ in most but not all elements of a model, are discussed in Section 16.4. Identification is important to establish if there is to be any hope of estimating $\widetilde{p}$ from observed data. A distribution $\widetilde{p}$ is said to be non-identified from $p(\mathbf{V})$ in a causal model if there exist two elements in the model which share $p(\mathbf{V})$ but differ in $\widetilde{p}$. No estimation strategy coherent for the entire model is possible for non-identified distributions.

In this chapter, rather than considering identification of counterfactual expectations, which were used to define causal effects and evaluate the quality of dynamic treatment regimes, we will concentrate on identification of distributions of potential outcomes, such as $p(Y(\mathbf{a}))$. Generally, causal effects and effects of dynamic treatment regimes are identified under the weaker single world model, while direct, indirect, and path-specific effects require the stronger multiple worlds model to yield identification, except in very simple cases.

16.2.4 Identification Of Causal Effects

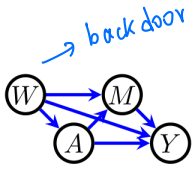
Under the single world model, for any value set $\mathbf{a}$ of $\mathbf{A} \subset \mathbf{V}$, the interventional distribution $p(\mathbf{V} \setminus \mathbf{A} \mid \mathrm{do}(\mathbf{a}))$ over counterfactuals $\{V(\mathbf{a}) \mid V \in \mathbf{V} \setminus \mathbf{A}\}$ is identified by

$$p(\mathbf{V} \setminus \mathbf{A} \mid \mathrm{do}(\mathbf{a})) = \prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V \mid \mathrm{pa}(V))|_{\mathbf{A}=\mathbf{a}}. \tag{16.8}$$

This equation is known as the *g-formula* [22], the *manipulated distribution* [31], or the *truncated factorization* [16].

An intuitive interpretation of (16.8) is as follows. In causal models of a DAG $\mathcal{G}$, the value of each variable V is determined by values of $\mathrm{pa}(V)$, and an exogenous source of noise ϵ_V, via a structural equation f_V. The conditional distribution induced by $\mathrm{pa}(V)$, ϵ_V and f_V that captures the variation of V as a function of values $\mathbf{w}$ of $\mathrm{pa}(V)$ is simply $p(V \mid \mathbf{w}_{\mathrm{pa}(V)}) = p(V \mid \mathrm{do}(\mathbf{w}))$. An intervention operation that sets variables $\mathbf{A}$ to $\mathbf{a}$ in particular implies that the value of any variable $A \in \mathbf{A}$ is now a constant, and no longer determined by either $\mathrm{pa}(A)$ or ϵ_A via f_A. Thus the overall distribution of the remaining variables can be obtained from the DAG factorization (16.1) by simply dropping all terms from the factorization that are no longer relevant post-intervention, namely terms of the form $p(A \mid \mathrm{pa}(A))$, and setting all remaining occurrences of elements in $\mathbf{A}$ to appropriate values in $\mathbf{a}$.

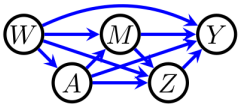
The g-formula has a number of well-known special cases. For instance, in Fig. 16.1 (a),



(a)

$$\begin{aligned} p(Y \mid \mathrm{do}(a)) &= \sum_{M,W} p(Y, M, W \mid \mathrm{do}(a)) \\ &= \sum_{M,W} p(Y \mid a, M, W) p(M \mid a, W) p(W) \\ &= \sum_{W} p(Y \mid a, W) p(W), \end{aligned}$$

which recovers the well-known adjustment or backdoor formula [16, 30]. In Fig. 16.1 (b),



(*b*)

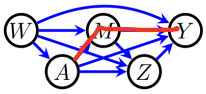
$$\begin{aligned} p(Y \mid \mathrm{do}(a,z)) &= \sum_{M,W,U} p(Y,M,W,U \mid \mathrm{do}(a,z)) \\ &= \sum_{M,W,U} p(Y \mid a,z,M,W,U)p(M \mid a,W,U)p(W) \\ &= \sum_{W,M} p(Y \mid a,z,M,W)p(M \mid a,W)p(W), \end{aligned}$$

marginalized out

which recovers the g-computation algorithm [22] for inferring causal effects in longitudinal studies (in this example for two time points, as that is the length shown in Fig. 16.1 (b)). Here and elsewhere, we use the summation notation for the sake of clarity of exposition, although summations only apply for discrete random variables. For continuous random variables similar expressions arise with an appropriately defined set of integrals.

16.2.5 Identification Of Path-Specific Effects

Distributions of nested counterfactuals involved in defining path-specific effects are more general objects than interventional distributions involved in defining average causal effects. As a consequence, not every such distribution is identified even in causal DAG models, and those that are identified require the stronger multiple worlds model. Identification for path-specific effects in a DAG is governed by a simple criterion known as the *recanting witness criterion* [1].



(*b*)

Let $\mathrm{ch}_{\mathcal{G}}(V)$ be the set $\{W \in \mathbf{V} \mid V \to W \text{ exists in } \mathcal{G}\}$. We omit $\mathcal{G}$ from the notation in cases where the graph is clear, yielding $\mathrm{ch}(V)$. If we are interested in the path-specific effect of $\mathbf{A}$ on $\mathbf{Y}$ along a set π of proper causal paths for $\mathbf{A}$ and $\mathbf{Y}$, we say a variable $W \in \mathrm{ch}(A)$ for some $A \in \mathbf{A}$ is a recanting witness for π if there exists a path in π with the first edge $A \to W$ and another proper causal path with the same first edge (but not necessarily the same final vertex in $\mathbf{Y}$) which is not in π. As an example, in Fig. 16.1 (b), if we are interested in the path-specific effect of A on Y along a single path $A \to M \to Y$, that is if $\pi = \{A \to M \to Y\}$, then M is a recanting witness for π, since the path $A \to M \to Z \to Y$ is a proper causal path for A and Y, is not an element of π, and has as its first edge $A \to M$ which is also the first edge of $A \to M \to Y$.

A well known result [1, 24] states that in a causal DAG $\mathcal{G}$, a path-specific from $\mathbf{A}$ to $\mathbf{Y}$ along a set of paths π is identified if and only if there does *not* exist a recanting witness for π. If the recanting witness does not exist, then the joint counterfactual distribution over variables $\{Y(\pi, \mathbf{a}, \mathbf{a}') \mid Y \in \mathbf{Y}\}$ is identified via a generalization of equation (16.8) called the *edge g-formula* [29]:

$$p(\{Y(\pi,\mathbf{a},\mathbf{a}')|Y \in \mathbf{Y}\}) = \sum_{\mathbf{V}\setminus(\mathbf{A}\cup\mathbf{Y})} \prod_{V\in\mathbf{V}\setminus\mathbf{A}} p(V|\mathbf{a}_{\mathrm{pa}^{\pi}(V)\cap\mathbf{A}}, \mathbf{a}'_{\mathrm{pa}^{\overline{\pi}}(V)\cap\mathbf{A}}, \mathrm{pa}(V)\setminus\mathbf{A}). \quad (16.9)$$

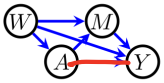
$$p(\mathbf{V}\setminus\mathbf{A} \mid \mathrm{do}(\mathbf{a})) = \prod_{V\in\mathbf{V}\setminus\mathbf{A}} p(V \mid \mathrm{pa}(V))|_{\mathbf{A}=\mathbf{a}}.$$

Just as the ordinary g-formula, the edge g-formula can be viewed as a truncated DAG factorization. However, in the edge g-formula a variable $A \in \mathbf{A} \cap \mathrm{pa}(V)$ in a Markov factor $p(V \mid \mathrm{pa}(V))$ can be set to either its value in $\mathbf{a}$ or its value in $\mathbf{a}'$, depending on whether the edge $A \to V$ is a part of a path in π or not.

As an example, a recanting witness does not exist for the path-specific effect of A on Y along the path set $\pi \equiv \{A \to M \to Y; A \to M \to Z \to Y\}$ in Fig. 16.1 (b). The counterfactual distribution $p(Y(\pi, a, a'))$ corresponding to this set of paths is then identified as

$$\sum_{W,M,Z} p(Y \mid M,W,Z,a)p(Z \mid M,W,a')p(M \mid W,a)p(W).$$

$\mathrm{pa}^{\pi}(Y)$ is the set of parents of Y along an edge which is a part of a path in π, and $\mathrm{pa}^{\overline{\pi}}(Y)$ is the set of all other parents of Y.



(a)

In the simple, but practically important case of Fig. 16.1 (a), where we are interested in the direct effect of A on Y, in other words in the path set $\pi \equiv \{A \to Y\}$, the counterfactual distribution $p(Y(\{A \to Y\}, a, a')) = p(Y(a, M(a')))$ is identified by the edge g-formula as

$$\sum_{W,M} p(Y \mid M, W, a) p(M \mid a', W) p(W).$$

If we are interested in using this counterfactual distribution to obtain the pure direct effect $\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a')]$, we recover the well-known *mediation formula* [17]:

$$\sum_{W,M} \{\mathbb{E}[Y \mid M, W, a] - \mathbb{E}[Y \mid M, W, a']\} p(M \mid a', W) p(W).$$

16.2.6 Identification Of Responses To Dynamic Treatment Regimes

Given a set of treatments $\mathbf{A}$ in a causal model represented by a DAG $\mathcal{G}$, the distribution $p(\{V(\mathbf{g_A}) | V \in \mathbf{V} \setminus \mathbf{A}\})$ over responses in $\mathbf{V} \setminus \mathbf{A}$ to a treatment regime $\mathbf{g_A}$ is identified via the following generalization of (16.8):

$$\prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V | \text{pa}(V) \setminus \mathbf{A}, \{A = g_A(\mathbf{W}_A) | A \in \text{pa}(V) \cap \mathbf{A}\}). \tag{16.10}$$

As an example, in Fig. 16.1 (b) the potential outcome $Y(\mathbf{g_A}) = Y(\{g_A, g_Z\})$, where g_A is a mapping from values of W to values of A, and g_Z is a mapping from values of M, W to values of Z, is identified as

$$\sum_{W,M} p(Y | A = g_A(W), Z = g_Z(M, W), M, W) p(M | A = g_A(W), W) p(W)$$

Equation (16.10) can be viewed as a version of (16.8) where values of $\mathbf{A}$ are not set to constants $\mathbf{a}$, but are instead set using appropriate functions in $\mathbf{g_A}$, which potentially creates dependence on other variables.

16.3 Causal Models Of A DAG With Hidden Variables

Identification results in the previous section assumed all variables in a DAG are observed. While this assumption allows an elegant characterization of identifiability via variations of the g-formula (16.8), (16.9), and (16.10), it is extremely unrealistic in practical applications. This motivates the study of causal models of a DAG where some variables are hidden. Unfortunately, the presence of hidden variables introduces a number of complications, both because hidden variables may prevent identification, and because functions $g(p(\mathbf{V}))$ of the observed distribution that correspond to counterfactual distributions that are identified can potentially become much more complex than (16.8), (16.9), and (16.10). In this section, we will describe a characterization of identifiable targets of causal inference in hidden variable causal DAGs, and identification algorithms that yield appropriate generalizations of the g-formula.

16.3.1 Latent Projections And Targets Of Inference

Identification in a causal model described by a hidden variable DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, where $\mathbf{V}$ are observed variables and $\mathbf{H}$ are hidden variables, is often described in terms of an acyclic directed mixed graph (ADMG) $\mathcal{G}(\mathbf{V})$ constructed from $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ via the latent projection operation described in Definition 2.3.2 from Chapter 2.

An ADMG is a mixed graph containing directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges that contains no directed cycles. In an ADMG, a pair of vertices can share at most two edges, and if the pair does share two edges, one of them is directed and one is bidirected. The latent projection ADMG represents an infinite class of hidden variable DAGs that, as we will describe later, all share identification theory.

Notation for a latent projection $\mathcal{G}(\mathbf{V})$ of a DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ with hidden variables $\mathbf{H}$ intentionally resembles the notation for marginalization in distributions. The latent projection operation can be viewed as a graphical analogue of the marginalization operation. A detailed exploration of this connection, and in particular a way of using mixed graphs derived from hidden variable DAGs to define statistical models that are supersets of marginals of distributions in DAG models is found in [9, 20].

Definitions of $Y(\mathbf{a})$ via (16.2), $Y(\pi, a, a')$ via (16.5), and $Y(\mathbf{g}_{\mathbf{A}})$ via (16.6) in a fully observed DAG $\mathcal{G}(\mathbf{V})$ carry over to a hidden variable DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ without change. Furthermore, if certain additional properties hold, these counterfactuals can be defined directly on $\mathcal{G}(\mathbf{V})_{\rightarrow}$, the edge subgraph of the latent projection $\mathcal{G}(\mathbf{V})$ of $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ that only contains directed edges.

These properties are listed below.

1. ∘ For $Y(\mathbf{a})$ in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, $Y \in \mathbf{V}$, and $\mathbf{A} \subseteq \mathbf{V}$.
2. ∘ For $Y(\mathbf{g}_{\mathbf{A}})$ in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, $Y \in \mathbf{V}$, $\mathbf{A} \subseteq \mathbf{V}$, and $\mathbf{W}_A \subseteq \mathbf{V}$ for every $A \in \mathbf{A}$.
3. ∘ For $Y(\pi, \mathbf{a}', \mathbf{a})$ in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, $Y \in \mathbf{V}$, $\mathbf{A} \subseteq \mathbf{V}$, and for any two directed paths π_i, π_j in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ such that the largest subsets of vertices in π_i, π_j that are in $\mathbf{V}$ form the same directed path $\widetilde{\pi}$ in $\mathcal{G}(\mathbf{V})$, either $\pi_i, \pi_j \in \pi$, or $\pi_i, \pi_j \notin \pi$.

If these properties hold, in any causal model of $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$,

1. ∘ Every $Y(\mathbf{a})$ defined using (16.2) in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ is equivalent to $Y(\mathbf{a})$ defined using (16.2) in $\mathcal{G}(\mathbf{V})_{\rightarrow}$.
2. ∘ Every $Y(\mathbf{g}_{\mathbf{A}})$ defined using (16.6) in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ is equivalent to $Y(\mathbf{g}_{\mathbf{A}})$ defined using (16.2) in $\mathcal{G}(\mathbf{V})_{\rightarrow}$.
3. ∘ Every $Y(\pi, \mathbf{a}', \mathbf{a})$ defined using (16.5) in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ is equivalent to $Y(\widetilde{\pi}, \mathbf{a}', \mathbf{a})$ defined using (16.2) in $\mathcal{G}(\mathbf{V})_{\rightarrow}$, where $\widetilde{\pi}$ is the set of directed paths formed from largest subsets of vertices in every element of π that intersect $\mathbf{V}$.

We will consider targets of inference defined on $\mathcal{G}(\mathbf{V})_{\rightarrow}$ directly, with the understanding that these are equivalent to targets of inference in any underlying hidden variable DAG model in the class represented by the latent projection $\mathcal{G}(\mathbf{V})$.

16.3.2 Conditional Mixed Graphs And Kernels

The identifying formulas described in Section 16.2 can be viewed as obtained by an algorithm that, in a single step, drops terms corresponding to elements in $\mathbf{A}$ from a DAG factorization, and possibly relabels remaining occurrences of $\mathbf{A}$ in remaining terms. In hidden variable DAGs represented by ADMGs, many counterfactual distributions of interest

are not identifiable. In addition, even for distributions that are identifiable, their identifying formulas are obtained from more complicated algorithms that perform sequences of operations that drop terms in a particular order. To effectively describe how these algorithms operate, we introduce a special type of graph and a special type of distribution that represent intermediate outputs these algorithms produce.

A *conditional ADMG (CADMG)* $\mathcal{G}(\mathbf{V}, \mathbf{W})$ is an ADMG with a vertex set $\mathbf{V} \cup \mathbf{W}$, where the set of vertices $\mathbf{W}$ are marked as *fixed*, and no edges of the form $\circ \rightarrow W$, or $\circ \leftrightarrow W$ exist. Vertices in $\mathbf{V}$ in a CADMG are meant to represent random variables as in an ordinary DAG, while vertices in $\mathbf{W}$ are meant to represent variables that were previously random, but which are currently set to a constant. Note that there are no restrictions on vertices in $\mathbf{V}$. In particular there may exist $V \in \mathbf{V}$ such that edges $\circ \rightarrow V$, or $\circ \leftrightarrow V$ also do not exist.

A bidirected path in a CADMG (or ADMG) is a path consisting entirely of bidirected edges. Given $V \in \mathbf{V}$, the maximum subset of $\mathbf{V}$ connected to V by a bidirected path (and V itself) is called the district [19] (or a c-component [34]) of V, abbreviated as $\text{dis}_{\mathcal{G}}(V)$. The set of districts in a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$ is denoted by $\mathcal{D}(\mathcal{G}(\mathbf{V}, \mathbf{W}))$ and forms a partition of elements in $\mathbf{V}$.

A *kernel* $q_{\mathbf{V}}(\mathbf{V} \mid \mathbf{W})$ is a mapping from values $\mathbf{w}$ of $\mathbf{W}$ to normalized densities $q_{\mathbf{V}}(\mathbf{V} \mid \mathbf{w})$ over $\mathbf{V}$. A conditional distribution is a type of kernel, though other types of kernels are possible. Given a subset $\mathbf{A} \subseteq \mathbf{V}$, marginalization and conditioning for kernels are defined in the usual way:

$$q_{\mathbf{V}}(\mathbf{A} \mid \mathbf{W}) \equiv \sum_{\mathbf{V} \setminus \mathbf{A}} q_{\mathbf{V}}(\mathbf{V} \mid \mathbf{W}); \qquad q_{\mathbf{V}}(\mathbf{V} \setminus \mathbf{A} \mid \mathbf{A} \cup \mathbf{W}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V} \mid \mathbf{W})}{q_{\mathbf{V}}(\mathbf{A} \mid \mathbf{W})}.$$

Kernels were called Q-factors in [34, 32]. The dependence of kernels on values of $\mathbf{W}$ was kept implicit in the Q-factor notation.

An ADMG $\mathcal{G}(\mathbf{V})$ representing a hidden variable DAG $\mathcal{G}(\mathbf{H} \cup \mathbf{V})$, and the corresponding observed data distribution $p(\mathbf{V})$ are special cases of a CADMG and a kernel, respectively. CADMGs and kernels involved in identification are derived from $\mathcal{G}(\mathbf{V})$ and $p(\mathbf{V})$ by sequential applications of the fixing operation, defined in [20].

16.3.3 The Fixing Operation

Given a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, a vertex $V \in \mathbf{V}$ is said to be fixable if $\text{de}_{\mathcal{G}}(V) \cap \text{dis}_{\mathcal{G}}(V) = \{V\}$, where $\text{de}_{\mathcal{G}}(V)$ (descendants of V) is the set of all vertices Z in $\mathcal{G}(\mathbf{V}, \mathbf{W})$, including V itself, with a directed path from V to Z. Given a fixable vertex V in $\mathcal{G}$ define the fixing operator on graphs $\phi_V(\mathcal{G}(\mathbf{V}, \mathbf{W}))$ to be an operator that produces a CADMG $\mathcal{G}(\mathbf{V} \setminus \{V\}, \mathbf{W} \cup \{V\})$ which is obtained from $\mathcal{G}(\mathbf{V}, \mathbf{W})$ by removing all edges pointing to V, and marking V as fixed.

Given a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, a kernel $q_{\mathbf{V}}(\mathbf{V} \mid \mathbf{W})$, and any fixable V in $\mathcal{G}$, define the fixing operator on kernels $\phi_V(q_{\mathbf{V}}(\mathbf{V} \mid \mathbf{W}); \mathcal{G}(\mathbf{V}, \mathbf{W}))$ to be one that produces the kernel

$$\widetilde{q}_{\mathbf{V} \setminus \{V\}}(\mathbf{V} \setminus \{V\} \mid \mathbf{W} \cup \{V\}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V} \mid \mathbf{W})}{q_{\mathbf{V}}(V \mid \text{nd}_{\mathcal{G}}(V) \cup \mathbf{W})},$$

where $\text{nd}_{\mathcal{G}}(V) \equiv (\mathbf{V} \cup \mathbf{W}) \setminus \text{de}_{\mathcal{G}}(V)$ is the set of non-descendants of V in $\mathcal{G}(\mathbf{V}, \mathbf{W})$.

A sequence $\langle V_1, \ldots, V_k \rangle$ is said to be fixable in $\mathcal{G}$ if V_1 is fixable in $\mathcal{G}$, V_2 is fixable in $\phi_{V_1}(\mathcal{G})$, V_3 is fixable in $\phi_{V_2}(\phi_{V_1}(\mathcal{G}))$, and so on. We extend the fixing operators for graphs and kernels in the natural way via function composition for any fixable sequence $\langle V_1, \ldots, V_k \rangle$:

$$\phi_{\langle V_1, \ldots, V_k \rangle}(\mathcal{G}) \equiv \phi_{V_k}(\ldots \phi_{V_2}(\phi_{V_1}(\mathcal{G})) \ldots)$$
$$\phi_{\langle V_1, \ldots, V_k \rangle}(q_{\mathbf{V}}; \mathcal{G}) \equiv \phi_{V_k}(\ldots \phi_{V_2}(\phi_{V_1}(q_{\mathbf{V}}; \mathcal{G}); \phi_{V_1}(\mathcal{G})) \ldots; \phi_{\langle V_1, \ldots, V_{k-1} \rangle}(\mathcal{G}))$$

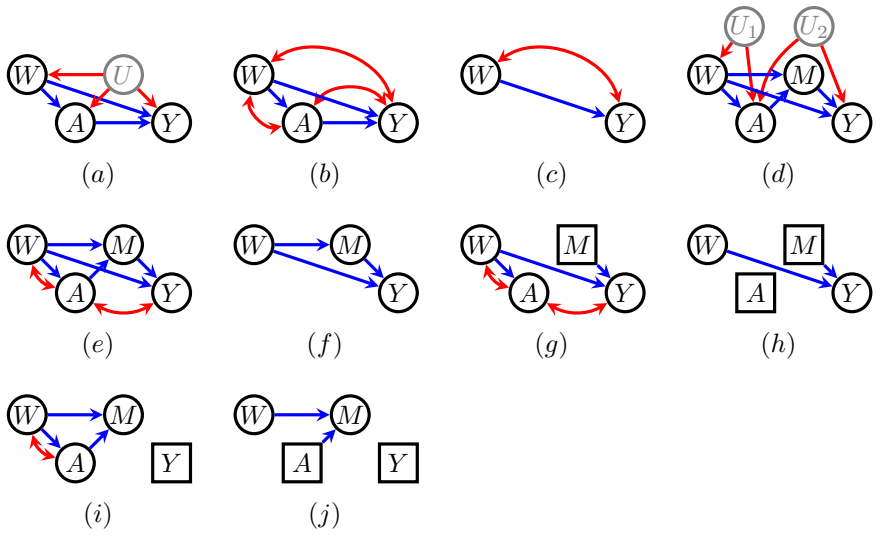


FIGURE 16.2: (a) A causal DAG with an unobserved common cause of the treatment A and the outcome Y which prevents identification of $p(Y(a))$. (b) The ADMG obtained from the DAG in (a) via the latent projection operation collapsing over the unobserved variable U. (c) A subgraph of the ADMG in (b) relevant for identification of $p(Y \mid \text{do}(a))$. (d) A causal DAG with an unobserved common cause of the baseline variables W, the treatment A and the outcome Y. This DAG also contains a mediator M that "captures" all the causal influence of A on Y that is also not confounded by U_2. (e) The ADMG $\mathcal{G}$ obtained from the DAG in (e) via the latent projection operation collapsing over the unobserved variables U_1, U_2. (f) A subgraph of the ADMG in (e) relevant for identification of $p(Y \mid \text{do}(a))$. (g) The CADMG $\phi_M(\mathcal{G})$ obtained from the ADMG in (e). (h) The CADMG $\phi_{\langle M,A\rangle}(\mathcal{G}) = \phi_{\{M,A\}}(\mathcal{G})$ obtained from the ADMG in (e). (i) The CADMG $\phi_Y(\mathcal{G})$ obtained from the ADMG in (e). (j) The CADMG $\phi_{\langle Y,A\rangle}(\mathcal{G}) = \phi_{\{Y,A\}}(\mathcal{G})$ obtained from the ADMG in (e).

Theorem 3.4.1 (Rules of *do* Calculus)
Let G be the directed acyclic graph associated with a causal model as defined in (3.2)*, and let $P(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z, and W, we have the following rules.*

Rule 1 (*Insertion/deletion of observations*):

$$P(y \mid \hat{x}, z, w) = P(y \mid \hat{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z) \mid X, W)_{G_{\overline{X}}}. \tag{3.31}$$

Rule 2 (*Action/observation exchange*):

$$P(y \mid \hat{x}, \hat{z}, w) = P(y \mid \hat{x}, z, w) \quad \text{if } (Y \perp\!\!\!\perp Z) \mid X, W)_{G_{\overline{X}\underline{Z}}}. \tag{3.32}$$

Rule 3 (*Insertion/deletion of actions*):

$$P(y \mid \hat{x}, \hat{z}, w) = P(y \mid \hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X},\overline{Z(W)}}}, \tag{3.33}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

It is known [20] that any two distinct fixable sequences of the same set of variables applied to both graphs and kernels yield the same object. This invariance property justifies considering fixable sets, which are any subsets of $\mathbf{V}$ with vertices that can be arranged in a fixable sequence, and fixing operators $\phi_{\mathbf{W}}(\mathcal{G})$, $\phi_{\mathbf{W}}(q_{\mathbf{V}}; \mathcal{G})$. These operators are defined on any fixable subset $\mathbf{W}$ of $\mathbf{V}$, and obtained via any fixable sequence on elements in $\mathbf{W}$.

16.3.4 The ID Algorithm

Given a hidden variable DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ representing a causal model, fix disjoint subsets $\mathbf{A}, \mathbf{Y}$ of $\mathbf{V}$ representing the set of treatments and outcomes, respectively, where we are interested in identification of the interventional distribution $p(\mathbf{Y} \mid \mathrm{do}(\mathbf{a}))$. Identification of these types of distributions in hidden variable DAGs was phrased in terms of a recursive algorithm called the ID algorithm [34, 28]. Here we give a simple reformulation of the ID algorithm in terms of the fixing operator ϕ, CADMGs and kernels, where the recursion is folded into the iterative application of the fixing operator.

Let $\mathcal{G}(\mathbf{V})$ be the latent projection of $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ onto observable vertices $\mathbf{V}$, and let $\mathbf{Y}^* \equiv \mathrm{an}_{\mathcal{G}(\mathbf{V})_{\mathbf{V}\setminus\mathbf{A}}}(\mathbf{Y})$, where for any subset $\mathbf{S}$ of $\mathbf{V}$, $\mathcal{G}(\mathbf{V})_{\mathbf{S}}$ is the subgraph of $\mathcal{G}(\mathbf{V})$ containing only vertices in $\mathbf{S}$ and edges in $\mathcal{G}(\mathbf{V})$ between pairs of vertices in $\mathbf{S}$. Then if for every $\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*})$, $\mathbf{V} \setminus \mathbf{D}$ is a fixable set,

$$p(\mathbf{Y} \mid \mathrm{do}(\mathbf{a})) = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*})} \phi_{\mathbf{V}\setminus\mathbf{D}}(p(\mathbf{V}); \mathcal{G}(\mathbf{V}))|_{\mathbf{A}=\mathbf{a}}. \tag{16.11}$$

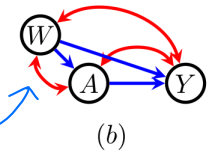
If some $\mathbf{V} \setminus \mathbf{D}$ is not fixable, then $p(\mathbf{Y} \mid \mathrm{do}(\mathbf{a}))$ is not identifiable, meaning that the ID algorithm is complete for identification of interventional distributions in non-parametric hidden variable DAG causal models. The proof of soundness of the original formulation of the ID algorithm appears in [34], and of completeness in [28]. The proof of soundness of the simplified version shown in (16.11) appears in [20]. Since preconditions for the application of (16.11), and all expressions within (16.11) itself were phrased in terms of the latent projection $\mathcal{G}(\mathbf{V})$, all causal models associated with a hidden variable DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ that yields the latent projection $\mathcal{G}(\mathbf{V})$ agree on which distributions $p(\mathbf{Y} \mid \mathrm{do}(\mathbf{a}))$ are identifiable and which are not, and further agree on all identifying functionals. It is for this reason that identification theory for counterfactuals in the presence of hidden variables is phrased in terms of the latent projection graph.

If $\mathbf{A} = \emptyset$, and $\mathbf{Y} = \mathbf{V}$, (16.11) can be viewed as a factorization of $p(\mathbf{V})$ known as the district or c-component factorization:

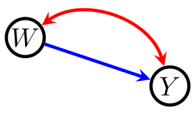
$$p(\mathbf{V}) = \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G})} q_{\mathbf{D}}(\mathbf{D} \mid \mathbf{V} \setminus \mathbf{D}) = \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G})} \phi_{\mathbf{V}\setminus\mathbf{D}}(p(\mathbf{V}); \mathcal{G}(\mathbf{V})). \tag{16.12}$$

This type of factorization has been used to define conditional independence supermodels of sets of marginals of hidden variable DAGs [10, 26, 20].

We now illustrate how (16.11) is applied with two examples. The first example is the graph in Fig. 16.2 (a), which contains an unobserved common cause of A and Y. The latent projection of this graph is the ADMG $\mathcal{G}(\{W, A, Y\})$ shown in Fig. 16.2 (b). Here $\mathbf{Y}^* = \mathrm{an}_{\mathcal{G}_{\{Y,W\}}}(Y) = \{Y, W\}$, with $\mathcal{G}_{\mathbf{Y}^*}$ shown in Fig. 16.2 (c). It's easy to verify that $\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*}) = \{\{W, Y\}\}$. Unfortunately, the set $\{A\}$ is not fixable, since Y is a descendant of A and lies in the district of A. From this we conclude that $p(Y \mid \mathrm{do}(a)) = p(Y(a))$ is not identified in the causal model represented by Fig. 16.2 (a).



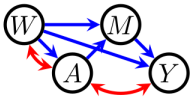
(b)



(c)

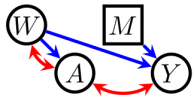
Consider now the graph in Fig. 16.2 (d). Like Fig. 16.2 (a), there is an unobserved common cause of A and Y, namely U_2. However, in Fig. 16.2 (d) there is, in addition, a mediator variable M which lies on a causal pathway from A to Y, indicated by the presence

of a directed path $A \to M \to Y$. In fact, this is the only directed path from A to Y, meaning that M captures or mediates all of the causal influence of A on Y. Finally, M is not a child of either U_1 or U_2, meaning it is determined entirely by W and A and remains unconfounded by hidden variables, unlike A and Y. The presence of a mediator of this type allows non-parametric identification of $p(Y \mid \mathrm{do}(a))$ in Fig. 16.2 (d), as was first noticed by [16]. We now give the identifying formula for $p(Y \mid \mathrm{do}(a))$ using (16.11).



(e)

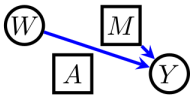
The latent projection of Fig. 16.2 (d) is the ADMG $\mathcal{G}(\{W, A, M, Y\})$ shown in Fig. 16.2 (e). Here $\mathbf{Y}^* = \mathrm{an}_{\mathcal{G}_{\{Y,W,M\}}}(Y) = \{Y, M, W\}$, with $\mathcal{G}_{\mathbf{Y}^*}$ shown in Fig. 16.2 (f). It's easy to verify that $\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*}) = \{\{Y\}, \{W\}, \{M\}\}$. In this case, the sets $\mathbf{V}\setminus\{Y\} = \{A, M, W\}$, $\mathbf{V}\setminus\{W\} = \{Y, M, A\}$, and $\mathbf{V}\setminus\{M\} = \{Y, W, A\}$ are fixable. We first consider $\{A, M, W\}$. M is fixable in Fig. 16.2 (e), which yields the CADMG in Fig. 16.2 (g), with the corresponding kernel



(g)

$$q_{\{W,A,Y\}}(W, A, Y|M) = \frac{p(W, A, M, Y)}{p(M|A, W)} = p(Y|M, A, W)p(A, W).$$

In this CADMG, A becomes fixable (it was not fixable in Fig. 16.2 (e)), yielding the CADMG in Fig. 16.2 (h), with the corresponding kernel



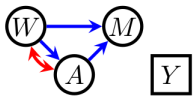
(h)

$$q_{\{W,Y\}}(W, Y|A, M) = \frac{q_{\{W,A,Y\}}(W, A, Y|M)}{q_{\{W,A,Y\}}(A|W, Y, M)} = \sum_A p(Y|M, A, W,)p(A, W).$$

Finally, in Fig. 16.2 (h), W is fixable, yielding a CADMG obtained from Fig. 16.2 (h) by drawing W as a square, with the corresponding kernel

$$q_{\{Y\}}(Y|W, A, M) = \frac{\sum_A p(Y|M, A, W,)p(A, W)}{\sum_{Y,A} p(Y|M, A, W,)p(A, W)} = \sum_A p(Y|M, A, W)p(A|W).$$

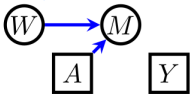
Similarly, the set $\{Y, W, A\}$ is also fixable. First, Y is fixable in Fig. 16.2 (e), yielding the CADMG in Fig. 16.2 (i), with the corresponding kernel



(i)

$$q_{\{W,A,M\}}(W, A, M|Y) = \frac{p(W, A, M, Y)}{p(Y|W, A, M)} = p(W, A, M).$$

Next, A becomes fixable in Fig. 16.2 (i), yielding the CADMG in Fig. 16.2 (j), with the corresponding kernel



(j)

$$q_{\{W,M\}}(W, M|A, Y) = \frac{q_{\{W,A,M\}}(W, A, M|Y)}{q_{\{W,A,M\}}(A|W, Y)} = p(M|A, W)p(W).$$

Finally, W is fixable in Fig. 16.2 (j), yielding a CADMG obtained from Fig. 16.2 (j) by drawing W as a square, with the corresponding kernel

$$q_{\{M\}}(M|W, A, Y) = \frac{q_{\{W,M\}}(W, M|A, Y)}{q_{\{W,M\}}(W|A, Y)} = p(M|A, W).$$

The set $\{Y, M, A\}$ is also fixable. We have already shown above that we can fix Y in Fig. 16.2 (e), and then fix A in Fig. 16.2 (i), yielding the graph in Fig. 16.2 (j), and the kernel $p(M|A, W)p(W)$. But M is fixable in Fig. 16.2 (j), yielding the kernel

$$q_{\{W\}}(W|M, A, Y) = \frac{p(M|A, W)p(W)}{p(M|A, W)} = p(W).$$

Combining these three kernels into the expression (16.11), where $\mathbf{Y} \setminus \mathbf{Y}^* = \{W, M\}$, and

$$p(\mathbf{Y} \mid \mathrm{do}(\mathbf{a})) = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*})} \phi_{\mathbf{V}\setminus\mathbf{D}}(p(\mathbf{V}); \mathcal{G}(\mathbf{V}))|_{\mathbf{A}=\mathbf{a}}.$$

evaluating $q_{\{M\}}(M \mid W, A, Y)$ at $A = a$ yields

$$\begin{aligned} p(Y \mid \mathrm{do}(a)) &= \sum_{W,M} \phi^a_{\{Y,W,A\}}(p; \mathcal{G})\, \phi_{\{A,M,W\}}(p; \mathcal{G})\, \phi_{\{Y,M,A\}}(p; \mathcal{G}) \\ &= \sum_{W,M} q_{\{M\}}(M|Y, W, a)\, q_{\{Y\}}(Y|A, M, W)\, q_{\{W\}}(W|Y, M, A) \\ &= \sum_{W,M} p(M|a, W) \left(\sum_A p(Y|M, A, W) p(A|W) \right) p(W). \end{aligned}$$

This is known as the front-door formula [16]. A connection between formulas of the above type, and the edge g-formula is given in [29].

In general, expressions obtained from (16.11) may become quite complicated. Nevertheless, (16.11) may be viewed as a generalization of the g-formula (16.8) appropriate to the hidden variable DAG setting. In particular, just as (16.8) is a truncated version of the DAG factorization (16.1), (16.11) is a truncated version of the district factorization in (16.12), in the sense that no kernel terms in (16.11) are densities over elements of $\mathbf{A}$. This is accomplished for each term in (16.11) by repeated applications of the fixing operation, which sequentially drop terms associated with variables in $\mathbf{V} \setminus \mathbf{D}$, which include the set $\mathbf{A}$ for every $\mathbf{D}$. Some fixing operations resemble conditioning, some resemble marginalization, and some resemble neither. Finally, just like in the g-formula (16.8), remaining terms in (16.11) that are functions of $\mathbf{A}$ are evaluated at $\mathbf{A} = \mathbf{a}$.

16.3.5 Controlled Direct Effects

An interesting special case of the identification problem described in the previous section occurs when $\mathbf{Y} = \{Y\}$, and $\mathbf{A} = \mathrm{pa}_{\mathcal{G}(\mathbf{V})}(Y)$ in a hidden variable DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$. The resulting counterfactual distribution $p(Y(\mathbf{a}))$ is used to defined *controlled direct effects* of the form

$$\mathbb{E}[Y \mid a, \mathbf{a}_{\mathrm{pa}_{\mathcal{G}(\mathbf{V})}(Y)\setminus\{A\}}] - \mathbb{E}[Y \mid a', \mathbf{a}_{\mathrm{pa}_{\mathcal{G}(\mathbf{V})}(Y)\setminus\{A\}}].$$

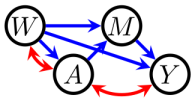
These types of effects are relevant in settings where we are interested in understanding the direct effect of A on Y within particular levels of all other observed direct causes of Y.

An ADMG $\mathcal{G}(\mathbf{V})$ is said to be an arborescence converging at Y if $\mathrm{an}_{\mathcal{G}}(Y) = \mathbf{V}$, and $\mathcal{D}(\mathcal{G}) = \{\mathbf{V}\}$. The distribution $p(Y(\mathbf{a}))$ of this type is identified if and only if the largest subset $\mathbf{V}'$ of $\mathbf{V}$ such that $\mathcal{G}_{\mathbf{V}'}$ is an arborescence converging at Y is $\{Y\}$. If $p(Y(\mathbf{a}))$ is identified, we have, as a special case of (16.11),

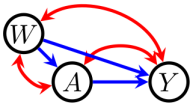
$$p(Y(\mathbf{a})) = q_Y(Y \mid \mathbf{V} \setminus \{Y\})|_{\mathbf{A}=\mathbf{a}} = \phi_{\mathbf{V}\setminus\{Y\}}(p(\mathbf{V}); \mathcal{G}(\mathbf{V}))|_{\mathbf{A}=\mathbf{a}}.$$

As an example, $p(Y(a))$ in Fig. 16.2 (b) is not identified since the graph in Fig. 16.2 (b) is an arborescence converging at Y. However, $p(Y(m, w))$ in $\mathcal{G}^{(e)}$ shown in Fig. 16.2 (e) is identified:

$$\begin{aligned} p(Y(m, w)) &= \phi_{\{W,M,A\}}(p(W, A, M, Y); \mathcal{G}^{(e)}) \\ &= \phi_{\{W,A\}}(p(Y|M, A, W)p(A, W); \phi_M(\mathcal{G}^{(e)}))|_{M=m} \\ &= \phi_{\{W\}} \left(\sum_A p(Y|M, A, W)p(A, W); \phi_{\{M,A\}}(\mathcal{G}^{(e)}) \right) \Big|_{M=m} \\ &= \sum_A p(Y|M, A, W)p(A|W)|_{M=m,W=w} = \sum_A p(Y|m, A, w)p(A|w). \end{aligned}$$



(e)



(b)

16.3.6 Conditional Causal Effects

A common variation of the problem of identification of joint interventional distributions considers instead identification of conditional interventional distributions of the form $p(Y_1(\mathbf{a}), \ldots, Y_k(\mathbf{a}) | Z_1(\mathbf{a}), \ldots, Z_m(\mathbf{a}))$, where $\mathbf{Y} = \{Y_1, \ldots, Y_k\}$, $\mathbf{Z} = \{Z_1, \ldots, Z_m\}$. This distribution is sometimes written as $p(\mathbf{Y}|\mathbf{Z}, \text{do}(\mathbf{a}))$. These types of distributions arise in causal inference applications where causal effects in particular subpopulations are of interest.

A simple modification of (16.11) gives a characterization of identifiability of these distributions. Let $\mathbf{Z}$ be partitioned into $\mathbf{W}$,$\mathbf{Z}'$ such that $\mathbf{W}$ is any maximal set with the property that for any $\mathbf{z}$, $p(\mathbf{Y} \mid \mathbf{z}, \text{do}(\mathbf{a})) = p(\mathbf{Y} \mid \mathbf{z}', \text{do}(\mathbf{w} \cup \mathbf{a}))$, and $\mathbf{z}'$, $\mathbf{w}$ are the appropriate partition of values in $\mathbf{z}$. Such a set $\mathbf{W}$, which is known to be unique and thus also maximum, can be obtained by application of the rule 2 of do-calculus [16]. Then $p(\mathbf{Y} \mid \mathbf{z}, \text{do}(\mathbf{a}))$ is identified from $p(\mathbf{V})$ if and only if $p(\mathbf{Y} \cup \mathbf{Z}' \mid \text{do}(\mathbf{w} \cup \mathbf{a}))$ is identified from $p(\mathbf{V})$, and is equal to

$$p(\mathbf{Y}|\mathbf{z}, \text{do}(\mathbf{a})) = \left. \frac{p(\mathbf{Y}, \mathbf{Z}'|\text{do}(\mathbf{w} \cup \mathbf{a}))}{p(\mathbf{Z}'|\text{do}(\mathbf{w} \cup \mathbf{a}))} \right|_{\mathbf{Z}'=\mathbf{z}'},$$

where $p(\mathbf{Y} \cup \mathbf{Z}'|\text{do}(\mathbf{w} \cup \mathbf{a}))$ is identified as usual via (16.11), and $p(\mathbf{Z}'|\text{do}(\mathbf{w} \cup \mathbf{a}))$ is obtained from $p(\mathbf{Y} \cup \mathbf{Z}'|\text{do}(\mathbf{w} \cup \mathbf{a}))$ via marginalization [27].

16.3.7 Path-Specific Effects

→ generalization

Path-specific effects in DAGs with all variables observed were not always identified due to the presence of recanting witnesses. Unsurprisingly, additional complications arise in hidden variable DAGs.

Fix $\mathbf{A}, \mathbf{Y}$ and a set of proper causal paths π for $\mathbf{A}$ and $\mathbf{Y}$, where each $A \in \mathbf{A}$ is the origin of at least one path in π. Let $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}}(\mathbf{Y})$. Then $\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*})$ is said to be a *recanting district* for π if there exists a path in π with the first edge of the form $A \to D$, where $A \in \mathbf{A}, D \in \mathbf{D}$, and another proper causal path not in π, with the first edge of the form $A \to D'$, $D' \in \mathbf{D}$.

In the absence of a recanting district in $\mathcal{G}(\mathbf{V})$, and if $p(\mathbf{Y} \mid \text{do}(\mathbf{a}))$ is identified from $p(\mathbf{V})$ in $\mathcal{G}(\mathbf{V})$, the joint counterfactual distribution over variables $\{Y(\pi, \mathbf{a}, \mathbf{a}') \mid Y \in \mathbf{Y}\}$ is identified as

$$p(\{Y(\pi, \mathbf{a}, \mathbf{a}') \mid Y \in \mathbf{Y}\}) = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V}\setminus\mathbf{D}}(p(\mathbf{V}); \mathcal{G}(\mathbf{V}))|_{\mathbf{A} \cap \text{pa}_{\mathcal{G}}(\mathbf{D}) = \tilde{\mathbf{a}}_{\mathbf{D}}}, \quad (16.13)$$

where $\tilde{\mathbf{a}}_{\mathbf{D}}$ is defined to be the subset of values of $\mathbf{a}$ corresponding to $\text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}$ if all elements in $\text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}$ are connected to elements in $\mathbf{D}$ via edges in π, defined to be the subset of values of $\mathbf{a}'$ corresponding to $\text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}$ if all elements in $\text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}$ are connected to elements in $\mathbf{D}$ via edges not in π, and defined to be the empty set if $\text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A} = \emptyset$. The absence of a recanting district guarantees these three possibilities are exhaustive.

If a recanting district is present, or $p(\mathbf{Y} \mid \text{do}(\mathbf{a}))$ is not identified, then $p(\{Y(\pi, \mathbf{a}, \mathbf{a}') \mid Y \in \mathbf{Y}\})$ is also not identified for some values $\mathbf{a}, \mathbf{a}'$. Just as (16.11) was an appropriate generalization of the g-formula (16.8) to hidden variable DAGs, so is (16.13) an appropriate generalization of the edge g-formula (16.9) to hidden variable DAGs.

As an example, consider Fig. 16.1 (c), where we are interested in the path-specific effect of A on Y via the path $A \to Z \to Y$. The latent projection of this graph is shown as a graph $\mathcal{G}^{(d)}$ in Fig. 16.1 (d). Here $\mathbf{Y}^* = \{W, M, Z, Y\}$, and $\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*}) = \{\{W\}, \{Z\}, \{M, Y\}\}$. Note that there is no recanting district – the district containing the first post-exposure variable on the only path of interest is $\{Z\}$, and no path other than $A \to Z \to Y$ has the first post-exposure variable in this district. Furthermore, $p(Y \mid \text{do}(a))$ is identifiable. Thus, the

counterfactual corresponding to the path-specific effect is identified:

$$p(Y(\pi, a, a')) = \sum_{W,Z,M} \left(\phi_{\{W,A,Z\}}(p; \mathcal{G}^{(d)})|_{A=a'}\right) \cdot \left(\phi_{\{W,A,M,Y\}}(p; \mathcal{G}^{(d)})|_{A=a}\right) \phi_{\{A,M,Z,Y\}}(p; \mathcal{G}^{(d)})$$
$$= \sum_{W,Z,M} \left(p(Y \mid a', M, Z, W)p(M \mid a', W)\right) p(Z \mid a, M, W)p(W).$$

On the other hand, if we were interested in the path-specific effect of A on Y along paths $\pi = \{A \to Z \to Y; A \to Y\}$, this path-specific effect is not identified. This is because the path $A \to M \to Y$ is not in π but has $A \to M$ as the first edge, while $A \to Y$ is a path in π. M and Y share a district in $\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*}$, where $\mathbf{Y}^* = \{W, M, Z, Y\}$. This implies $\{M, Y\}$ is a recanting district, and will prevent identification of $Y(\{A \to Z \to Y; A \to Y\}, a, a')$.

16.3.8 Responses To Dynamic Treatment Regimes

generalization

A general algorithm for identification of distributions $p(\{Y(\mathbf{g_A}) \mid Y \in \mathbf{Y}\})$ in causal models represented by a hidden variable DAG $\mathcal{G}(\mathbf{H} \cup \mathbf{V})$, and the corresponding latent projection $\mathcal{G}(\mathbf{V})$ was given in [32]. Here we reformulate this algorithm in terms of the fixing operator. Define the graph $\mathcal{G}(\mathbf{V})_{\mathbf{g_A}}$ to be an ADMG obtained from $\mathcal{G}(\mathbf{V})$ by removing all edges pointing into $\mathbf{A}$ and adding a directed edge $W \to A$ for any $W \in \mathbf{W}_A$, and the set $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}(\mathbf{V})_{\mathbf{g_A}}}(\mathbf{Y}) \setminus \mathbf{A}$. Then $p(\{Y(\mathbf{g_A}) \mid Y \in \mathbf{Y}\})$ is identified if $p(\mathbf{Y}^* \mid \text{do}(\mathbf{a}))$ is identified. Moreover, the identification formula is

$$p(\{Y(\mathbf{g_A}) \mid Y \in \mathbf{Y}\}) = \sum_{(\mathbf{Y}^* \cup \mathbf{A}) \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}); \mathcal{G}(\mathbf{V}))|_{\{A = g_A(\mathbf{W}_A) | A \in \mathbf{A}\}}.$$

The sum over $\mathbf{A}$ is vacuous if $\mathbf{g_A}$ is a set of deterministic policies, since in this case there is no variation in values of $\mathbf{A}$ in any $Y(\mathbf{g_A})$. This algorithm is conjectured, but currently not known, to be complete for identification of responses to dynamic treatment regimes in hidden variable DAGs models.

As an example, in Fig. 16.1 (c), if $\mathbf{g_A} = \{g_A(W), g_Z(M, W)\}$, $\mathcal{G}(\mathbf{V})$ is shown in Fig. 16.1 (d), and $\mathcal{G}(\mathbf{V})_{\mathbf{g_A}}$ is the same graph as $\mathcal{G}(\mathbf{V})$. Since $p(Y, M, W \mid \text{do}(a, z))$ is identified as

$$\phi_{\{A,M,Z,Y\}}(p; \mathcal{G}(\mathbf{V}))\phi_{\{W,A,Z\}}(p; \mathcal{G}(\mathbf{V}))_{A=a,Z=z},$$

which is equal to $p(W)p(Y \mid z, M, a, W)p(M \mid a, W)$, $p(Y(\mathbf{g_A}))$ is also identified as

$$\sum_{W,M,A,Z} \phi_{\{A,M,Z,Y\}}(p; \mathcal{G}(\mathbf{V}))\phi_{\{W,A,Z\}}(p; \mathcal{G}(\mathbf{V}))|_{\{A=g_A(W), Z=g_Z(M,W)\}}$$
$$= \sum_{W,M,A,Z} p(W)p(Y|Z = g_Z(M, W), M, A = g_A(W), W)p(M|A = g_A(W), W).$$

16.4 Linear Structural Equation Models

An important parametric causal model representable directly by an ADMG (and not necessarily a DAG) is the linear structural equation model (SEM). The linear SEM associated with an ADMG $\mathcal{G}(\mathbf{V})$ represents potential outcomes $V_i(\mathbf{v}_{\text{pa}_{\mathcal{G}}(V_i)})$ via linear functions

$f_{V_i}(\mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V_i)}, \epsilon_{V_i})) \equiv \sum_{M \in \mathrm{pa}_{\mathcal{G}}(V)} w_M M + \epsilon_V$, where $(\epsilon_{V_1}, \ldots, \epsilon_{V_k})$ is a random vector following the multivariate normal distribution $\mathcal{N}(0, \Omega)$, where Ω is a positive-definite matrix with the property that it contains a 0 entry in the ij cell for any V_i, V_j such that $V_i \neq V_j$ and $V_i \leftrightarrow V_j$ is absent in $\mathcal{G}(\mathbf{V})$. A well known result states that $p(\mathbf{V})$ specified via a linear SEM associated with an ADMG is multivariate normal. The value of the first moment of this distribution does not affect identification considerations, and is often assumed without loss of generality to be the 0 vector of the appropriate size.

Linear SEMs have a long pedigree in applied data analysis problems, particularly in the social sciences [35, 13]. Aside from its non-parametric generalizations given by the single world model and the multiple worlds model, current literature has generalized linear SEMs in a number of useful directions, such as factor analysis models [14] which which allow inclusion of meaningful unobserved variables, and non-Gaussian additive noise models [18] which permit inferences about causal directionality of edges from observed data. However, the classical linear SEM remains a very important and useful model class. An attractive property of this model is its relative simplicity for computation and statistical inference, and its additional algebraic structure layered on top of the non-parametric structure of the graphical causal model. This additional structure permits identification in a number of cases where identification is impossible non-parametrically.

Parameters of interest in linear SEMs are the coefficients which suffice to specify linear functions f_{V_i} for every $V_i \in \mathbf{V}$. Most interesting causal targets of inference can be phrased in terms of these coefficients. For example, if $A \in \mathrm{pa}_{\mathcal{G}}(Y)$, and the linear structural equation f_Y for Y is

$$Y = w_A A + \sum_{V \in \mathrm{pa}_{\mathcal{G}}(V) \setminus \{A\}} w_V V + \epsilon_Y,$$

then a simple substitution argument, and linearity of f_Y implies that w_A corresponds to the controlled direct effect:

$$\mathbb{E}[Y(A = 1, \mathbf{a})] - \mathbb{E}[Y(A = 0, \mathbf{a})] = w_A, \tag{16.14}$$

for any assignment $\mathbf{a}$ to $\mathrm{pa}_{\mathcal{G}}(Y) \setminus \{A\}$.

Identification results in linear SEMs take two forms, *global* and *generic*. A parameter w_V is said to be globally identified in a graph $\mathcal{G}(\mathbf{V})$, if it is a function of $p(\mathbf{V})$ in any linear SEM corresponding to $\mathcal{G}(\mathbf{V})$. A parameter is said to be generically identified in $\mathcal{G}(\mathbf{V})$ if it is a function of $p(\mathbf{V})$ in almost every element of the linear SEM corresponding to $\mathcal{G}(\mathbf{V})$, except perhaps a set of elements of measure zero. If $\mathcal{G}(\mathbf{V})$ is a DAG, then every parameter is globally identified due to (16.8) and (16.14).

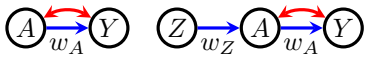
In linear SEMs associated with ADMGs, some parameters are generically, but not globally identified, and some are not identified at all. For example, in the linear SEM corresponding to Fig. 16.3 (a), given by

$$\begin{aligned} A &= \epsilon_A \\ Y &= w_A A + \epsilon_Y, \end{aligned}$$

where $\mathrm{cov}[\epsilon_A, \epsilon_Y] \neq 0$, the coefficient w_A, representing the average causal effect of A on Y is not identified from $p(\mathbf{V})$.

On the other hand, in Fig. 16.3 (b) known as the *instrumental variable graph*, corresponding to the model

$$\begin{aligned} Z &= \epsilon_Z \\ A &= w_Z Z + \epsilon_A \\ Y &= w_A A + \epsilon_Y, \end{aligned} \tag{16.15}$$



(a) (b)

FIGURE 16.3: (a) A simple ADMG representing a linear SEM where the coefficient w_A is not identified from $p(A, Y)$. (b) A more complex ADMG, the instrumental variable graph, representing a linear SEM where the coefficient w_A is generically, but not globally identified from $p(A, Y)$.

where $\text{cov}[\epsilon_A, \epsilon_Y] \neq 0$, the parameter w_A is not globally identified (since setting $w_Z = 0$ recovers the model in Fig. 16.3 (a)), but is generically identified as $w_A = \text{cov}[Z, Y]/\text{cov}[Z, A]$.

16.4.1 Global Identification Of Linear SEMs

An elegant characterization of global identifiability of the linear SEM model (that is global identifiability of every coefficient of every structural equation) was given in [8]. Specifically the linear SEM model associated with an ADMG $\mathcal{G}(\mathbf{V})$ is globally identified if and only if for every $V \in \mathbf{V}$, the largest subgraph of $\mathcal{G}(\mathbf{V})$ that forms an arborescence converging at V contains a single vertex.

There is a connection between this result and the earlier results on identifiability of controlled direct effects described in Section 16.3.5. Identification of controlled direct effects of A on Y, for any A and Y in an ADMG $\mathcal{G}(\mathbf{V})$ representing a causal model of a hidden variable DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ is characterized by the absence of an arborescence of size 2 or more converging on Y. This implies that if the largest convergent arborescence in $\mathcal{G}(\mathbf{V})$ has size 1, then every controlled direct effect is identified non-parametrically. In fact, this identification result applies, without change, to a linear SEM defined directly on an ADMG $\mathcal{G}(\mathbf{V})$, and implies the entire model is identified due to (16.14). The converse is also true, although this is considerably more difficult to show [8].

16.4.2 Generic Identification Of Linear SEMs

The characterization of generic identifiability of a linear SEM remains an open problem, despite decades of work [3, 2, 33, 12, 11, 7, 5]. It is known via computer algebra methods [12] that showing whether a particular parameter is identified in a linear SEM is a decidable problem, although its computational complexity is not currently known. We describe a general method based on features of the graph described in [11], although this method is known to not be complete, and further extensions are possible.

In an ADMG $\mathcal{G}(\mathbf{V})$ a half-trek between vertices V_i, V_j is a path of one of the following two types:

$$\pi \equiv \underbrace{V_i \to \circ \to \ldots \to \circ \to V_j}_{\text{Right}(\pi)}$$

$$\pi \equiv \underbrace{V_i}_{\text{Left}(\pi)} \leftrightarrow \underbrace{\circ \to \ldots \to \circ \to V_j}_{\text{Right}(\pi)}$$

with subset $\text{Left}(\pi)$ defined to be $\emptyset$ for the first type of half-trek, and $\text{Left}(\pi)$ and $\text{Right}(\pi)$ defined as shown otherwise. A half-trek can be an empty path (a path with no edges and $V_i = V_j$). A set of half-treks is called a *system* if no two half-treks in the set share initial or

final vertices. A system of half-treks $\{\pi_1, \ldots, \pi_k\}$ is said to have *no sided intersection* if for every $i, j \in \{1, \ldots, k\}$, and $i \neq j$,

$$\text{Left}(\pi_i) \cap \text{Left}(\pi_j) = \emptyset = \text{Right}(\pi_i) \cap \text{Right}(\pi_j).$$

In Fig. 16.4 (a), $A \to B \to C$ and $A \leftrightarrow D \to E$ are half-treks while $A \leftrightarrow E \leftrightarrow D$ and $A \to B \leftrightarrow C$ are not. The pair of half-treks $\pi_1 \equiv C \to D \to E$ and $\pi_2 \equiv D \leftrightarrow A$ form a system of half-treks $\{\pi_1, \pi_2\}$ from $\{C, D\}$ to $\{A, E\}$. This system has no sided intersection, since $D \in \text{Right}(\pi_1) \cap \text{Left}(\pi_2)$. On the other hand, π_1 and $\pi_3 \equiv A \leftrightarrow D$ form a system of half-treks $\{\pi_1, \pi_3\}$ from $\{A, C\}$ to $\{D, E\}$ with a sided intersection since $D \in \text{Right}(\pi_1) \cap \text{Right}(\pi_3)$. This illustrates that initial and final vertices in a half-trek matter for the purposes of determining the existence of sided intersections.

For any $V \in \mathbf{V}$ in an ADMG $\mathcal{G}(\mathbf{V})$ denote the set of siblings of V, $\text{sib}_{\mathcal{G}(\mathbf{V})}(V) \equiv \{W \leftrightarrow V \mid W \in \mathbf{V}\}$. For every $V \in \mathbf{V}$ define the following subset of vertices reachable from V by a half-trek:

$$\text{htr}_{\mathcal{G}(\mathbf{V})}(V) \equiv \left\{ W \in \mathbf{V} \setminus (\{V\} \cup \text{sib}_{\mathcal{G}(\mathbf{V})}(V)) \;\middle|\; \begin{array}{c} V \leftrightarrow \circ \to \ldots \to \circ \to W \\ \text{or} \\ V \to \circ \to \ldots \to \circ \to W \end{array} \right\}.$$

A (possibly empty) vertex set $\mathbf{Y} \subseteq \mathbf{V}$ satisfies the *half-trek criterion* with respect to $V \in \mathbf{V}$ if $|\mathbf{Y}| = |\text{pa}_{\mathcal{G}(\mathbf{V})}(V)|$, $\mathbf{Y} \cap (\{V\} \cup \text{sib}_{\mathcal{G}(\mathbf{V})}(V)) = \emptyset$, and there is a system of half-treks with no sided intersection from $\mathbf{Y}$ to $\text{pa}_{\mathcal{G}(\mathbf{V})}(V)$. If $\text{pa}_{\mathcal{G}(\mathbf{V}}(V) = \emptyset$, then $\mathbf{Y} = \emptyset$ satisfies the half-trek criterion with respect to V.

The half-trek criterion gives a very general condition for generic identifiability of the linear SEM called **HTC-identifiability.** A linear SEM model represented by an ADMG $\mathcal{G}(\mathbf{V})$ is HTC-identifiable if there is a set

$$\{\mathbf{Y}_V \subseteq \mathbf{V} \mid \mathbf{Y}_V \text{ satisfies the half-trek criterion with respect to } V\}$$

in $\mathcal{G}(\mathbf{V})$, and there is a total ordering $\prec$ on $\mathbf{V}$ in $\mathcal{G}(\mathbf{V})$ such that $W \prec V$ whenever $W \in \mathbf{Y}_V \cap \text{htr}_{\mathcal{G}(\mathbf{V}}(V)$. HTC-identifiability guarantees that, except in a small set of elements of the model, every coefficient of every structural equation is a rational function of the observed distribution $p(\mathbf{V})$. These functions are obtained by solving systems of linear equations following the ordering $\prec$, with the details given in [11].

A closely related condition called **HTC-nonidentifiability** strongly precludes identification of the model. A linear SEM model represented by an ADMG $\mathcal{G}(\mathbf{V})$ is HTC-nonidentifiable if for every set of sets $\{\mathbf{Y}_V \subseteq \mathbf{V} \mid V \in \mathbf{V}\}$ in $\mathcal{G}(\mathbf{V})$, either some $\mathbf{Y}_V$ does not satisfy the half-trek criterion with respect to V or there exist $\mathbf{Y}_V$, $\mathbf{Y}_W$ in this set such that $V \in \mathbf{Y}_W$ and $W \in \mathbf{Y}_V$. HTC-nonidentifiability guarantees the existence of an infinite set of elements of a linear SEM model corresponding to $\mathcal{G}(\mathbf{V})$ that agree on the observed distribution $p(\mathbf{V})$.

The graph $\mathcal{G}^{(a)}$ in Fig. 16.4 (a) is HTC-identifiable. Let

$$Y_A = \emptyset; \quad Y_B = \{E\}; \quad Y_C = \{B\}; \quad Y_D = \{B\}; \quad Y_E = \{C\}.$$

Then each set above satisfies the half-trek criterion with respect to its vertex:

- Y_A: since $\text{pa}_{\mathcal{G}^{(a)}}(A) = \emptyset$;
- Y_B: since $Y_B \cap (\{B\} \cup \text{sib}_{\mathcal{G}^{(a)}}(B)) = \{E\} \cap \{B, A\} = \emptyset$, and the single half-trek system $\{E \leftrightarrow A\}$ has no sided intersection;
- Y_C: since $Y_C \cap (\{C\} \cup \text{sib}_{\mathcal{G}^{(a)}}(C)) = \{B\} \cap \{C, D\} = \emptyset$, and the single half-trek system consisting of the empty half-trek $\{\emptyset\}$ has no sided intersection;

- Y_D: since $Y_D \cap (\{D\} \cup \text{sib}_{\mathcal{G}^{(a)}}(D)) = \{B\} \cap \{A, C, D, E\} = \emptyset$, and the single half-trek system $\{B \to C\}$ has no sided intersection;
- Y_E: since $Y_E \cap (\{E\} \cup \text{sib}_{\mathcal{G}^{(a)}}(E)) = \{C\} \cap \{A, D, E\} = \emptyset$, and the single half-trek system $\{C \to D\}$ has no sided intersection.

Finally, we have

$$\begin{array}{cc} \text{htr}_{\mathcal{G}^{(a)}}(A) = \{C\}; & \text{htr}_{\mathcal{G}^{(a)}}(A) \cap Y_A = \emptyset \\ \text{htr}_{\mathcal{G}^{(a)}}(B) = \{C, D, E\}; & \text{htr}_{\mathcal{G}^{(a)}}(B) \cap Y_B = \{E\} \\ \text{htr}_{\mathcal{G}^{(a)}}(C) = \{E\}; & \text{htr}_{\mathcal{G}^{(a)}}(C) \cap Y_C = \emptyset \\ \text{htr}_{\mathcal{G}^{(a)}}(D) = \{B\}; & \text{htr}_{\mathcal{G}^{(a)}}(D) \cap Y_D = \{B\} \\ \text{htr}_{\mathcal{G}^{(a)}}(E) = \{B, C\}; & \text{htr}_{\mathcal{G}^{(a)}}(E) \cap Y_E = \{C\} \end{array}$$

Thus, any ordering $\prec$ which asserts $C \prec E \prec B \prec D$ will satisfy the criterion for HTC-identifiability. Both HTC-identifiability and HTC-nonidentifiability are properties that can be checked in polynomial time using algorithms for determining maximum flow in a graph. The details are found in [11].

HTC-identifiability and HTC-nonidentifiability are not complete methods in the sense that there exist identifiable and non-identifiable models that are neither HTC-identifiable nor HTC-nonidentifiable. An interesting example occurs in Fig. 16.4 (b). The linear SEM corresponding to this graph is neither HTC-identifiable, nor HTC-nonidentifiable. However, it is possible to show that the linear SEM corresponding to this graph is identified by first applying (16.12), which yields

$$\begin{aligned} p(A, B, C, D, E) &= q_{\{A,B,C\}}(A, B, C \mid D, E) q_{\{D,E\}}(D, E \mid A, B, C) \\ &= [p(A, C \mid B, D, E) p(B \mid D)] \cdot [p(E \mid B, D) p(D)], \end{aligned}$$

with CADMGs corresponding to $q_{\{A,B,C\}}$ and $q_{\{D,E\}}$ shown in Fig. 16.4 (c) and (d), respectively.

The linear SEM corresponding to Fig. 16.4 (d) satisfies the HTC-identifiability criterion because Fig. 16.4 (d) is a simple graph (every vertex pair shares at most one edge). In any simple graph $\mathcal{G}$, for every vertex V, $\text{pa}_{\mathcal{G}}(V)$ satisfies the half-trek criterion for V and any ordering topological for $\mathcal{G}$ will then satisfy HTC-identifiability. See also proposition 1 in [11].

The linear SEM corresponding to Fig. 16.4 (c) is HTC-identifiable with

$$\mathbf{Y}_D = \mathbf{Y}_E = \emptyset; \quad \mathbf{Y}_B = \{A\}; \quad \mathbf{Y}_C = \{D, E\}; \quad \mathbf{Y}_A = \{C, D\},$$

and any ordering $\prec$ where $C \prec A$. This implies that every parameter of $q_{\{A,B,C\}}$ and $q_{\{D,E\}}$ is generically identified, and thus so is every parameter in $p(A, B, C, D, E)$ in a linear SEM for the graph in Fig. 16.4 (b). General methods that combine the half-trek criterion and district decomposition appear in [6, 5, 7].

16.5 Summary

In classical statistical inference parameter identification is often trivial, and thus not usually discussed. In causal inference problems, parameter identification is in general a subtle problem.

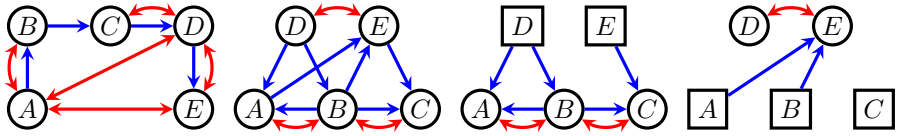


FIGURE 16.4: (a) An ADMG. (b) An ADMG representing a linear SEM that is not generically identified according to the half-trek criterion. (c) (d) CADMGs corresponding to the district factorization of the observed distribution corresponding to the linear SEM in (a), where each district factor is identified according to the half-trek criterion.

In causal models represented by directed acyclic graphs (DAGs) where all relevant variables are observed, counterfactual responses to interventions that set variables to constants or via a known function of other variables are identified by versions of the g-formula (16.8), (16.10). Counterfactual responses corresponding to path-specific effects are identified by the edge g-formula (16.9) if and only if no recanting witness variables exist in the DAG.

In hidden variable DAGs, counterfactual responses to interventions may no longer be identified. However, complete identification algorithms exist for counterfactual distributions used to define causal effects, conditional causal effects, controlled direct effects, and path-specific effects. These algorithms, originally given a recursive formulation in [34, 28, 27], have been given a simple formulation using the fixing operator defined in [20], and latent projection mixed graphs representing classes of hidden variable DAGs that share identification theory. In addition, very general identification results are known for counterfactual responses to variables being set according to a known policy, and in parametric causal models defined on acyclic directed mixed graphs (ADMGs) termed linear structural equation models.

16.6 Acknowledgments

Some definitions, examples and existing identification results that appear in this chapter also appear in [25].

Bibliography

[1] Chen Avin, Ilya Shpitser, and Judea 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.

[2] C. Brito and J. Pearl. Graphical condition for identification in recursive sem. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 47–54. AUAI Press, Arlington, VA, 2006.

[3] Carlos Brito and Judea Pearl. A graphical criterion for the identification of causal

effects in linear models. In *Eighteenth national conference on Artificial intelligence*, pages 533–538. American Association for Artificial Intelligence, 2002.

[4] Bibhas Chakraborty and Erica E. M. Moodie. *Statistical Methods for Dynamic Treatment Regimes (Reinforcement Learning, Causal Inference, and Personalized Medicine)*. Springer, New York, 2013.

[5] Bryant Chen. Identification and overidentification of linear structural equation models. In *Advances in Neural Information Processing Systems*, volume 29, pages 1579–1587. Curran Associates, Inc., 2016.

[6] Bryant Chen, Jin Tian, and Judea Pearl. Testable implications of linear structural equation models. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2424–2430, 2014.

[7] M. Drton and L. Weihs. Generic identifiability of linear structural equation models by ancestor decomposition. *Scandinavian Journal of Statistics*, 2016.

[8] Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *Annals of Statistics*, 39(2):865–886, 2011.

[9] Robin J. Evans and Thomas S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, pages 1–30, 2014.

[10] Robin J. Evans and Thomas S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. Unpublished, 2015.

[11] Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *Annals of Statistics*, 40(3):1682–1713, 2012.

[12] L. D. Garcia-Puente, S. Spielvogel, and S. Sullivant. Identifying causal effects with computer algebra. In *Proceedings of the Twenty-sixth Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2010.

[13] Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12, 1943.

[14] R. B. Kline. *Principles and Practice of Structural Equation Modeling*. The Guilford Press, 2005.

[15] Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 411–420. Morgan Kaufmann, San Francisco, 2001.

[16] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

[17] Judea Pearl. The causal mediation formula – a guide to the assessment of pathways and mechanisms. Technical Report R-379, Cognitive Systems Laboratory, University of California, Los Angeles, 2011.

[18] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Scholkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, pages 2009–2053, 2014.

[19] Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30:962–1030, 2002.

[20] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. Working paper, 2017.

[21] Thomas S. Richardson and Jamie 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.

[22] James M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[23] James M. Robins and Sander Greenland. Identifiability and exchangeability of direct and indirect effects. *Epidemiology*, 3:143–155, 1992.

[24] Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science (Rumelhart special issue)*, 37:1011–1035, 2013.

[25] Ilya Shpitser. Identification in causal models with hidden variables. *Journal of the French Statistical Society (to appear)*, 2017.

[26] Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

[27] Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 437–444. AUAI Press, Corvallis, Oregon, 2006.

[28] Ilya Shpitser and Judea 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.

[29] Ilya Shpitser and Eric Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433–2466, 2016.

[30] Ilya Shpitser, Tyler VanderWeele, and James M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 527–536. AUAI Press, 2010.

[31] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, New York, 2 edition, 2001.

[32] Jin 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.

[33] Jin Tian. Parameter identification in a class of linear structural equation models. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1970–1975. AAAI Press, Palo Alto, CA, 2009.

[34] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, volume 18, pages 519–527. AUAI Press, Corvallis, Oregon, 2002.

[35] Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.