15

Causal Concepts and Graphical Models

Vanessa Didelez
Leibniz Institute for Prevention Research and Epidemiology & Department of Mathematics, University of Bremen, Germany

CONTENTS

15.1 Introduction

The notion of causality has been much examined, discussed and debated, in science and philosophy, over many centuries (see accounts in [67, 49, 36]). In the social discourse, it often refers to aspects of responsibility, such as moral questions ('whose fault is it?') or historical questions ('what caused the second world war?'). However, for the purpose of the present chapter, we focus mainly on statistical contexts where causal inference uses data to inform *decisions* about *actions*, for example with a view to public health policies. In particular we consider different approaches to formalizing causal enquiries which, despite

subtle differences, all build on a probabilistic graphical representation of the problem at hand. We have deliberately chosen to present different approaches in order to illustrate the flexibility and representational power of graphical models.

Overview

We start by clarifying the notation and terminology used in this chapter as well as the role of graphical models, recalling how they relate to conditional independence structures. In Section 15.2, a brief tour through different causal notations and frameworks introduces common key concepts as well as differences between formal approaches. Two ways of combining these with graphical models are addressed and compared in Section 15.3. A central question of causal inference is whether a desired causal effect is *identified* from observational data. Early ideas of how to obtain graph-based answers are discussed in Section 15.4.

The usefulness of graphs becomes especially clear when investigating structural sources of bias. They allow to detect, for example, different ways how confounding may be introduced or removed from an analysis; see Section 15.5. We conclude with an outlook onto the many further topics of causal inference where graphs play a facilitating role.

Notation and Terminology

We use $\mathbf{X}_V = \mathbf{X} = (X_1, \ldots, X_K), V = \{1, \ldots, K\}$, to refer to a random vector in connection with a graph $\mathcal{G} = (V, E)$. Within specific examples we alternatively use X, Y, Z, C, T, W, U etc. as random variables. The domains of random variables are given by $\mathcal{X}_k = \{x_k \in \mathbb{R} : p(x_k) > 0\}$, $k \in V$, and similar for other variables. We loosely refer to p as distribution or probability, but let it also stand for either the probability density or probability mass function of a distribution. Occasionally we will write $\mathbf{X}^i$ or Y^i etc. when referring to an individual in the sample or population, but mostly we suppress this additional index. For $A \subset V$, subvectors and induced subgraphs are denoted by $\mathbf{X}_A$ and $\mathcal{G}_A$. Graphical parents, children, descendants and non-descendants in a directed acyclic graph (DAG) are denoted by $\mathrm{pa}(A), \mathrm{ch}(A), \mathrm{de}(A), \mathrm{nd}(A)$ respectively (see formal definitions in Section 1.6 of Chapter 1).

The Role of Graphs

The standard problem of causal inference in statistics is concerned with analyzing observational data in order to infer, and often quantify, causal relations. We would intuitively expect that the validity of such inference requires specialized methods. In particular it relies on assumptions that are somewhat different in their nature, and arguably stronger, than those for traditional statistical inference. Only a thorough understanding of these assumptions, and indeed of the causal target of inference itself, enables us to come up with ways of checking them either empirically or based on subject matter plausibility. Such an understanding is therefore a prerequisite for assessing the soundness of a proposed causal conclusions in any given situation. This is of great importance as causal findings are designed to inform decisions and policy interventions with practical consequences. For example, it can have very different implications if we claim that dietary fat intake causes coronary heart disease than if we say that dietary fat intake is only associated with coronary heart disease.

Graphical approaches assist us with formalizing causal targets of inference. Moreover, they have proved especially useful for being transparent and general about the assumptions required for causal conclusions. Hence they facilitate the detection and elimination of possible sources of bias, or suggest specific sensitivity analyses. Also, we often gain a better insight into the logic behind methods of causal inference when using graphs. Finally, graphs are sometimes a causal target of inference in their own right.

Before going into the details of how to combine causal concepts and graphical models, let us first revisit some key aspects of the latter and establish that the *Markov properties do not automatically supply directed edges with any causal meaning*. We mainly consider DAGs, but see Section 15.6 for references to approaches that use other types of graphs.

Recall that the distribution for a random vector $\mathbf{X} = (X_1, \ldots, X_K)$ factorizes according to a DAG $\mathcal{G} = (V, E)$ if the joint distribution satisfies

$$p(\mathbf{x}) = \prod_{k \in V} p(x_k | \mathbf{x}_{\mathrm{pa}(k)}). \tag{15.1}$$

Conditional independencies can be read off the DAG using various criteria. For example, the above factorization is equivalent to $X_k \perp\!\!\!\perp \mathbf{X}_{\mathrm{nd}(k) \setminus \mathrm{pa}(k)} \,|\, \mathbf{X}_{\mathrm{pa}(k)}$. Most commonly, the d-separation criterion based on blocking of paths can be used (or equivalently the moralization criterion); see Section 1.8 in Chapter 1 or also [40]. For example, the two DAGs $1 \longleftarrow 2 \longleftarrow 3$ and $1 \longrightarrow 2 \longrightarrow 3$ imply the exact same single conditional independence $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ because node 2 d-separates nodes 1 and 3 in both DAGs. When graphs with different orientations of edges represent the same set of conditional independencies they are called *Markov equivalent*. But even if a DAG is uniquely determined by its Markov properties it still represents no more than certain conditional independencies.

Hence, if we want a graph to encode some notion of causality an extra ingredient is required. Graphical and causal modeling can, in fact, be combined in many ways resulting in differing representational power. We aim to provide the reader with a broad overview and will focus on essentially two approaches: (i) augmenting traditional conditional independence DAGs with a separate type of non-random nodes representing interventions; (ii) retaining the original set of nodes pertaining to the domain variables but modifying the meaning of edges, hence supplying a causal interpretation on top of the graphical Markov properties. We return to this distinction in more detail in Section 15.3.

15.2 Association versus Causation: Seeing versus Doing

Traditional statistical regression models for an outcome Y and an exposure X specify some aspect of the conditional distribution $p(y|x)$, or in words: the distribution of Y when we *observe* $X = x$. This describes an association in the sense of how *seeing* different values of X helps us to *predict* possibly different (expected) values of Y. A regression-based analysis would typically be accompanied by a warning that 'association is not causation'. The warning seems necessary so that people do not expect the predicted change in Y to actually materialize when they *intervene* in a system to *manipulate* the value of X. In other words, the warning is given so that individuals or policy makers do not base *decisions* about their *actions* on a finding that is 'just associational'. Intervening can be understood as actively *doing* something to the system instead of passively observing it. To make this explicit, Pearl [49] introduced the so-called $\mathrm{do}(\cdot)$-notation defined below.

As a toy example, consider the positive association between the amount of books in a household and the household income. This presumably reflects that higher education is accompanied by more books as well as better paid jobs and that with a higher income one can afford more books. It would be misguided to think that one's income can be raised purely by increasing the number of books in one's home.

If we want to address these issues formally we require a notation allowing us to express

for instance that, while there can be an association between X and Y, a manipulation of X may not necessarily result in a corresponding change in Y. We can regard this as making an explicit distinction between concepts of *association* on the one hand and *causation* on the other. Note, however, that none of the approaches covered below actually define causality. They are simply all based on the premise that causation, in contrast to association, is relevant to decisions about actions, such as interventions or manipulations. Pragmatically, without going into philosophical subtleties, we say that X is causal for Y if some manipulation of X has an effect on Y.

The following brief overview of some notations will give an idea of the key concepts employed in the causal literature, each providing a slightly different angle on the problem (see also overviews [21, 55, 13]). Section 15.3 revisits these in more detail and combines them with graphical models.

Do–Calculus.

An intuitive notation distinguishing conditioning on observing versus intervening has been introduced by Pearl [49]. The former is denoted by $p(y; \text{see}(X = \widetilde{x}))$ and can be equated with ordinary conditioning $p(y; \text{see}(X = \widetilde{x})) = p(y|\widetilde{x})$. Here, 'seeing' refers to X taking its natural value, where 'natural' is always relative to a given context. For example, 'natural' could be the situation where in an observational study we simply ask individuals of a specific population how many books they have at home. In contrast, the notation $p(y; \text{do}(X = \widetilde{x}))$[1] refers to the distribution of Y under the situation where X has been *forced* to take the value $\widetilde{x}$ by some *intervention*. For example, the state could provide every household with $\widetilde{x}$ books. With $Y =$ 'income' and $X =$ '# books', it is intuitively plausible that $p(y|\widetilde{x}) \neq p(y; \text{do}(X = \widetilde{x}))$. This inequality is a formal way to state that association is not causation. Note that several different notations have been used to indicate conditioning on intervention, see [46, 47, 67, 49, 40, 11].

Pearl [47, 49] further develops a do-calculus which relates the corresponding intervention to a DAG and then uses graphical rules to convert conditioning on 'doing' into conditioning on 'seeing'. The defining property of a *causal DAG* demands that a $\text{do}(X_j = \widetilde{x}_j)$-intervention modularly replaces the factor $p(x_j|\mathbf{x}_{\text{pa}(j)})$ by $\mathbf{1}\{x_j = \widetilde{x}_j\}$ in the factorized joint distribution (15.1) [33]. This do-intervention is called an *atomic* intervention as it affects a single variable that is being set to a specific value [46].

Regime Indicator.

We can consider the above 'seeing' and 'doing' as two types of *regimes*, a natural one and a set of interventional ones. Here, regimes refer to external circumstances under which we expect some aspects of the joint distribution of $\mathbf{X}$ to differ [13]. In many situations a do-intervention is somewhat idealized or hypothetical. How would one for instance fix the dietary fat intake or BMI of a person exactly at a given value? It is also implicit in the notation that the manner in which a variable is manipulated is irrelevant. However, in practice it may matter whether a medical treatment is, for example, given orally or as an injection. For greater generality, we may therefore want to consider a possibly larger and more detailed set $\mathcal{S}$ of different regimes describing different circumstances under which a system might be observed and manipulated. Each regime then induces a different probability measure for the joint distribution of $\mathbf{X}$. Hence, let σ be an indicator for the regime taking values in $\mathcal{S}$ and index the distribution accordingly, so that $p(\mathbf{x}; \sigma = s) = p(\mathbf{x}; s)$, $s \in$

[1] Pearl [49] uses the notation $p(y|\,\text{do}(X = \widetilde{x}))$; but here we want to avoid confusion with the measure theoretic definition of probabilistic conditioning on random variables. Instead we regard $\text{do}(X = \widetilde{x})$ as an index to the distribution, $p(\cdot; \text{do}(X = \widetilde{x}))$.

$\mathcal{S}$ denotes the joint distribution of $\mathbf{X}$ under regime s[2]. As addressed in Section 15.3.1, suitable factorizations of the joint density under each regime express assumptions about interventions, for example, what type of interventions in which variables take place and how these then affect further variables.

Returning to the example of an outcome Y and exposure X, the joint distributions under different regimes are given by $p(y, x; s), s \in \mathcal{S}$. Under any regime we always have $p(y, x; s) = p(x; s)p(y|x; s)$. If a particular regime $s \in \mathcal{S}$ indicates an intervention that physically manipulates X we can express this exactly by specifying $p(x; \sigma = s)$. For instance, let the above 'seeing' be denoted by $\sigma = \emptyset$, also called the 'idle' or 'observational' regime[3]. In contrast, forcing X to be $\widetilde{x}$ is denoted $\sigma = \widetilde{x}$, $\widetilde{x} \in \mathcal{X}$, such that in this case $\mathcal{S} = \text{'}\emptyset\text{'} \cup \mathcal{X}$, where $\mathcal{X}$ is the domain of X. Hence $p(x; \sigma = \emptyset)$ is the distribution of the exposure we observe naturally, while $p(x; \sigma = \widetilde{x}) = \mathbf{1}\{x = \widetilde{x}\}$ is an atomic intervention. With these choices of regimes, $p(y; \sigma = \emptyset)$ denotes the distribution of Y when X arises naturally, and $p(y|\widetilde{x}; \sigma = \emptyset)$ is the corresponding conditional distribution of Y when we passively observe $X = \widetilde{x}$. In contrast, $p(y; \sigma = \widetilde{x})$ is the distribution of Y when X is forced to take on the value $\widetilde{x}$; now $p(y|x; \sigma = \widetilde{x})$ is undefined unless $x = \widetilde{x}$, and $p(y|\widetilde{x}; \sigma = \widetilde{x}) = p(y; \sigma = \widetilde{x})$ due to the particular choice of $p(x; \sigma = \widetilde{x})$ as a point-distribution.

As with the do$(\cdot)$-notation, we can express that 'association is not causation' by allowing $p(y|\widetilde{x}; \sigma = \emptyset)$ to differ from $p(y; \sigma = \widetilde{x})$. But we can now also formulate different types of interventions, for example experimental settings where X is randomly generated from a distribution $\widetilde{p}_X$. To represent this, let $\mathcal{S}$ denote the set of interventions defined by demanding that $p(x; \sigma = \widetilde{p}_X) = \widetilde{p}_X(x), \widetilde{p}_X \in \mathcal{S}$. Now, $p(y; \sigma = \widetilde{p}_X)$ describes the behavior of the outcome Y when X is drawn from the distribution $\widetilde{p}_X$. Interventions that do not fix a variable at a value, but just 'nudge' it, adding a random error, or somehow shift its distribution are sometimes called *soft* interventions [22].

Another important type of regimes is given by conditional or dynamic interventions [56]. Here, we may want to force X to take on a value that is a specified function g_X of pre-exposure covariates C. The purpose is to reflect, for example, a treatment strategy that is adapted to the patient's history. Under such a regime we have that $p(x|C = c; \sigma = g_X) = \mathbf{1}\{x = g_X(c)\}$. More generally even, we may also include situations where potentially more aspects of the system than just a single variable X are affected by a manipulation, possibly even in a partially unknown manner.

Aspects that remain the same under different regimes are called *stable* or *invariant*. These properties can formally be expressed by statements such as $Y \perp\!\!\!\perp \sigma \,|\, X$, meaning $p(y|x; \sigma = s) = p(y|x; \sigma = s'), s \neq s' \in \mathcal{S}$. Such invariances can be read off suitable graphs via d-separations if these graphs include a node for σ. These graphs are called *intervention graphs* and are addressed in Section 15.3.1.

Potential Outcomes.

A third notation in the context of causal inference uses *potential outcomes* [61, 62]. If, as above, we want to consider some causal effect of X on an outcome Y, we define the potential outcome $Y(\widetilde{x})$ to be the value of Y that we would observe if X were set (forced) to $\widetilde{x}$. Hence this approach is essentially based on atomic interventions. Similar to before, the possibility that $p(Y(x) = y)$ does not equal $p(y|x)$ allows us to express that causation is not association. But note that potential outcomes are formulated at the level of variables

[2]Several authors have used the notation F_X to indicate an atomic intervention in X in a similar way to our use of σ here, see [46, 47, 40, 11, 13]. We use σ for greater generality, alluding to intervention strategies as in [14].

[3]We use the terms 'seeing', observational / natural / unmanipulated / idle regime interchangeably — they all refer to the distribution from which data is actually available which can sometimes be a randomized controlled trial.

instead of distributions. They can hence be used to express individual causal effects as functions of $Y^i(\widetilde{x})$ for different $\widetilde{x}$. Potential outcomes also allow us to express 'cross-world' independencies such as $Y(\widetilde{x}) \perp\!\!\!\perp W(x') \,|\, (Z, X)$ for two different values $\widetilde{x}, x'$.

These examples illustrates why potential outcomes are also referred to as *counterfactuals*. As X^i can only ever take one value at a time for a given individual i, only one of $Y^i(\widetilde{x}), \widetilde{x} \in \mathcal{X}$, can ever be observed and the others are 'counter to the fact'. One may therefore want to be careful with implicit or explicit assumptions about the joint distribution of counterfactuals as these can never be observed together [10]. Robins and Richardson [55] advocate a restriction to models and assumptions allowing only 'single-world' interventions.

It may not be immediately obvious how potential outcomes can be combined with graphs, and often they are not. An overview of different approaches to combine them is given in [55]. One approach uses structural equation models (SEMs). These assume that all variables are functionally related at an individual level in a way that is invariant to interventions. The functional relations correspond to the equations of an SEM. the equations, in turn, relate each variable to its parents in an associated DAG. Finally, the functional relations induce a joint distribution on all counterfactuals. We return to this in more detail in Section 15.3.2.

Causal Effects

As alluded to earlier, we say that X has a causal effect on Y if an intervention in the former affects the distribution of the latter. For example, the presence of a causal effect means that for two values $\widetilde{x} \neq \widetilde{x}'$ we have $p(y; \mathrm{do}(X = \widetilde{x})) \neq p(y; \mathrm{do}(X = \widetilde{x}'))$. Similarly, if $s, s' \in \mathcal{S}$ denote two different, not necessarily atomic, interventions in X, then a causal effect means that $p(y; \sigma = s) \neq p(y; \sigma = s')$. Particular causal parameters can be formalized as functions of intervention distributions. For instance, the average causal effect (ACE) is given by $E(Y; \mathrm{do}(X = \widetilde{x})) - E(Y; \mathrm{do}(X = \widetilde{x}'))$, but other contrasts such as odds ratios or risk ratios are popular. Using potential outcomes the individual causal effect (ICE) is defined as $Y^i(\widetilde{x}) - Y^i(\widetilde{x}')$ while the ACE is the expectation over the individuals in the population $E(Y(\widetilde{x}) - Y(\widetilde{x}'))$. In the following we typically refer to the whole intervention distributions as causal effects unless stated otherwise.

15.3 Extending Graphical Models for Causal Reasoning

As announced earlier, we now consider two different approaches to extending DAGs with a view to a causal interpretation. The first adds a particular type of node representing experimental manipulation or intervention. The second approach modifies the semantics of a DAG to obtain a *causal DAG*.

15.3.1 Intervention Graphs

Intervention graphs are conditional independence graphs augmented by a separate type of node for σ as an indicator for the regime[4]. As statistical models they retain the interpretation in terms of Markov properties. However, this raises a question as the regime indicator is not a random variable. It may not make sense to specify a distribution over the possible

[4]Unfortunately, the terminology is not in agreement here. Lauritzen [40] calls these 'intervention graphs', while Dawid [11] calls them 'augmented' and uses 'intervention' graphs for causal DAGs; they can also be regarded as 'decision' or 'influence diagrams' [39]. As graphs can be augmented in several different ways, we revert to 'intervention graphs'.

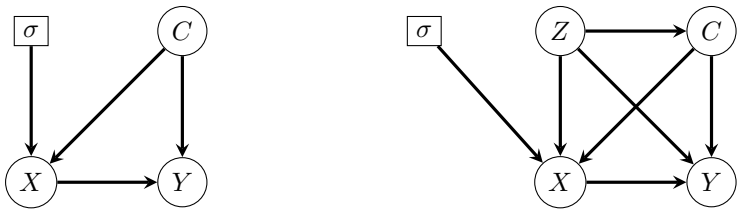

FIGURE 15.1: Examples of intervention DAGs.

regimes, especially if some of them correspond to 'what if' scenarios such as: 'banning smoking' or 'what if doctors followed strategy 1 versus strategy 2 when treating HIV patients' etc. Nevertheless we can make conditional independence statements such as $\mathbf{X}_B \perp\!\!\!\perp \sigma \,|\, \mathbf{X}_A$ to indicate that the conditional distribution of $\mathbf{X}_B$ given $\mathbf{X}_A$ is identical across the particular choice of regimes $\mathcal{S}$ (for a formal treatment see Dawid and Constantinou [6]). In fact, this *stability* [14] or *invariance* [33, 53] of conditional distributions across regimes is central to causal inference where a key aim is to predict effects of interventions from observational data. Clearly, in order to be able to infer anything about the former based on the latter at least some aspects of the two situations must be assumed 'similar'. Graphs are useful to make such assumptions explicit. We explain this with an example.

Example: Consider the case where a patient suffers from some ailment and may or may not take a treatment. Let X be a treatment indicator, C a blood-measurement and Y a health outcome. The different regimes are: (P) the patient decides what treatment to take, (O) the doctor decides the treatment based on old guidelines, or (N) the doctor decides the treatment based on new guidelines. A typical question would be how to compare regime O versus N based on data gathered only under regime P and whether this is possible at all. (We return to this as a question of *identifiability* later.) As the circumstances are not necessarily the same in the three situations we may want our model to allow three different joint distributions $p(\cdot; \sigma = s), s = \text{P, O, N}$. For each we have the factorization

$$p(x, c, y; s) = p(y|x, c; s)p(x|c; s)p(c; s).$$

Some of $p(x|c; s)$ may be unknown, for example $p(x|c; \sigma = \text{P})$, while others are determined by background knowledge, for example the old guidelines may specify a decision rule according to which treatment is given if C is larger than a threshold τ, yielding $p(X = 1|c; \sigma = \text{O}) = \mathbf{1}\{c > \tau\}$. Now assume that there is reason to believe that, across the three regimes, only the conditional distribution of treatment $p(x|c; s)$ is possibly different, but the conditional distribution of Y given (X, C) as well as the marginal distribution of C remain the same. Formally, this is implied by the equalities $p(y|x, c; s) = p(y|x, c; s')$ and similarly $p(c; s) = p(c; s'), s \neq s'$. The assumptions are represented in Figure 15.1(a), where the box around σ reminds us that this node is not a random variable but a regime indicator. It is easy to justify the invariance of the marginal distribution of C if the covariates in C are measured before 'deciding' the value of X. Its distribution will then obviously be the same under all regimes. A sufficient condition for the invariance of Y given (X, C) across regimes, in this example, would be that the only information available to the decision maker, in any regime, is the blood measurement. However, under regime P this may be implausible.

The ideas sketched above can be employed very flexibly to represent a wide range of

situations and assumptions. An intervention graph, as defined next, describes the probabilistic dependence structure of a system that factorizes according to a given DAG, where certain factors are determined by interventions.

Definition 15.3.1 (Intervention DAG and Model). *Consider the random vector* $\mathbf{X} = (X_1, \ldots, X_K)$*, on vertices* $V = \{1, \ldots, K\}$ *of a DAG* $\mathcal{G}$*. Let* $\mathcal{S}$ *denote a set of regimes and let* $p(\cdot; \sigma = s), s \in \mathcal{S}$*, be the joint distributions under the respective regimes. The augmented DAG* $\mathcal{G}^\sigma = (V \cup \{\sigma\}, E^\sigma)$ *is called the* intervention DAG *for* $\mathbf{X}$ *under regimes* $\mathcal{S}$ *if it has the following properties:*
(i) the node σ *is a source node (has no incoming directed edges),*
(ii) each distribution $p(\mathbf{x}; s), s \in \mathcal{S}$*, factorizes according to* $\mathcal{G}$*,*
(iii) for disjoint $A, B \subset V$*, whenever* B *is d-separated from* σ *by* A *in* $\mathcal{G}^\sigma$ *we have:* $p(\mathbf{x}_B|\mathbf{x}_A; s) = p(\mathbf{x}_B|\mathbf{x}_A; s')$*, for all* $s \neq s'$*. This is denoted by* $\mathbf{X}_B \perp\!\!\!\perp \sigma|\mathbf{X}_A$*.*

A key aspect of Intervention DAGs is that the factors $p(x_k|\mathbf{x}_{\mathrm{pa}(k)}; \sigma = s)$ for $k \in \mathrm{ch}(\sigma)$ characterize the particular regime s, while the remaining factors are assumed invariant. Moreover, the graph itself is assumed invariant; with the above definition it is, for instance, not possible to describe a situation where under one regime we have $j \longleftarrow k$ and under the other regime we have $j \longrightarrow k$.

Example ctd.: This example illustrates that a given intervention DAG can be implausible if important information is ignored. First, note that the sub-DAG on $\{C, X, Y\}$ in Figure 15.1(a) is Markov equivalent to any other orientation of the edges. This means that as a conditional independence DAG it imposes no restrictions at all on the joint distribution of (C, X, Y). However, adding $\sigma \longrightarrow X$, as shown, means that there is no other Markov equivalent DAG on $\{C, X, Y, \sigma\}$. The stability assumptions therefore make this a unique model that can in principle be *verified with data from more than one regime.* For instance if we have three data sets, one each under P, O, and N, respectively, then it can be checked empirically that the marginal of C and the conditional of Y given (C, X) are the same in all data sets.

In practice, we rarely have data sets from more than one regime (but see [53] for examples with data from different regimes). Assumptions such as in Figure 15.1(a) then need to be justified with subject matter knowledge about, for example, the physical mechanisms linking (C, X, Y) and the contemplated intervention(s). In many situations it will be plausible that, if the system is specified in sufficient detail, certain aspects of it will remain invariant across a range of specified situations, and the edges out of σ then correspond to specific physical manipulations. As a counterexample assume that an important factor Z has been omitted, for example the patient's temperature which could be influencing the treatment decision of the patient or doctor. Then the DAG of Figure 15.1(a) becomes implausible and we may want to consider the DAG in (b) instead, where $Y \perp\!\!\!\perp \sigma \,|\, (Z, C, X)$ but $Y \not\perp\!\!\!\perp \sigma \,|\, (C, X)$, hence the incorrectness of (a).

The appropriateness of a particular intervention DAG also depends on the set of regimes $\mathcal{S}$ considered. Figure 15.1(a) may be suitable to represent $\sigma \in \mathcal{S} = \{\mathrm{O}, \mathrm{N}\}$ where X is assigned based on the old or new guidelines if these use only the blood measurement as information. But it may not be suitable when including the patient as decision maker, so when $\sigma \in \mathcal{S}' = \{\mathrm{P}, \mathrm{O}, \mathrm{N}\}$, as the patient may have measured her temperature before deciding on treatment.

15.3.2 Causal DAGs

Causal DAGs could be defined as intervention DAGs where the set of regimes consist of the natural (observational) regime and atomic interventions in *all* nodes [13]. Under these two

conventions the indicator σ and the set $\mathcal{S}$ can be dropped, as σ would simply have edges into all nodes and all interventions are of the same type. Pearl's $\mathrm{do}(\cdot)$-notation is often combined with causal DAGs. However, the notion of a causal DAG is in fact more in the spirit of Spirtes et al. [67] who refer to *unmanipulated* and *manipulated* populations. Pearl [47], in contrast, links causal DAGs, or *causal diagrams*, very much with with SEMs and potential outcomes.

The following defines causal DAGs such that atomic interventions in arbitrary nodes are adequately represented by a modular modification to the factorized joint distribution.

Definition 15.3.2 (Causal DAG). *Consider a DAG $\mathcal{G} = (V, E)$ and a random vector $\mathbf{X} = (X_1, \ldots, X_K)$ with distribution p. Then $\mathcal{G}$ is called a* causal DAG *for $\mathbf{X}$ if p satisfies the following:*
(i) p factorizes, and thus is Markov, according to $\mathcal{G}$, and
(ii) for any $A \subset V$ and any $\widetilde{\mathbf{x}}_A, \mathbf{x}_B$ in the domains of $\mathbf{X}_A, \mathbf{X}_B$, where $B = V \backslash A$,

$$p(\mathbf{x}; \mathrm{do}(\widetilde{\mathbf{x}}_A)) = \prod_{k \in B} p(x_k | x_{\mathrm{pa}(k)}) \prod_{j \in A} \mathbf{1}\{x_j = \widetilde{x}_j\}. \tag{15.2}$$

The second factor of (15.2) makes explicit that $\mathrm{do}(\widetilde{\mathbf{x}}_A)$ refers to interventions that fix $\mathbf{X}_A$ at the given value $\widetilde{\mathbf{x}}_A$. In other words, the difference between the observational distribution $p(\mathbf{x})$ and the intervention distribution $p(\mathbf{x}; \mathrm{do}(\widetilde{\mathbf{x}}_A))$ is that all factors $p(x_j | x_{\mathrm{pa}(j)}), j \in A$, are removed and replaced by degenerate probabilities $\mathbf{1}\{x_j = \widetilde{x}_j\}$, while all remaining factors $p(x_k | x_{\mathrm{pa}(k)}), k \in B$, stay the same. Equation (15.2) is therefore known as the *truncated factorization* and the substitution of individual factors $p(x_j | x_{\mathrm{pa}(j)})$ as *causal modularity* [33].

The validity of (15.2) in a given context could in principle be decided by carrying out all possible atomic interventions. It is more common to employ subject matter knowledge, and include unobservables, to justify a causal DAG. Similar to the points made in connection with intervention graphs, the plausibility of causal modularity will crucially depend on the set of nodes (measured or unmeasured) being sufficiently rich. This is often expressed in the recommendation that when a causal DAG is posited, the absence of possible further edges and further nodes needs to be justified. In other words, for a DAG to be causal it needs to include all 'common causes' of its nodes [36].

Example ctd.: To compare intervention graphs with causal DAGs consider again Figure 15.1. The corresponding causal DAGs would simply omit the node σ, but would additionally assume that atomic interventions in C and Z are also correctly represented. This makes a difference in (b), where the direction of the edge between Z and C is not relevant to the assumptions expressed in the intervention DAG because a reversal of the edge is Markov equivalent to the original intervention DAG. However, it is relevant in the causal DAG as it implies a different causal ordering between Z and C.

Randomisation. It can easily be seen that (15.2) would be obtained as the ordinary conditional distribution of $\mathbf{X}_B$ given $\mathbf{X}_A = \widetilde{\mathbf{x}}_A$ in a graph where all incoming edges into $\mathbf{X}_A$ have been deleted. Such a truncated DAG corresponds to the situation where $\mathbf{X}_A$ was drawn randomly and independently of any predecessors while all other relations remain the same. In other words, it mimics the ideal randomized experiment reflecting the special status of randomized trials for causal inference.

Similarly it can be seen that (15.2) is obtained from the original joint distribution upon dividing by $p(x_j | x_{\mathrm{pa}(j)})$, $j \in A$. This motivates the principle of 'inverse probability of treatment weighting' (IPTW). With IPTW a dataset is reweighted to recreate empirically

the situation of a randomly assigned treatment or unconfounded exposure [59, 36]. Hence, IPTW empirically 'removes' the arrows into the variable targeted by an intervention.

Faithfulness

With Definition 15.3.2, the absence of a directed edge from X_j to X_k in a causal DAG can be interpreted as 'X_j has no direct causal effect on X_k relative to the given set of nodes'. More precisely, if we were to fix the parent nodes of X_k, and then manipulate X_j, $j \notin \mathrm{pa}(j)$, we would see no change in the distribution of X_k. Conversely, one might say that a directed edge always represents a direct causal effect. However, this is not implied by the Definition 15.3.2. It requires the additional assumption of *causal faithfulness* [67].

To make this more formal, recall that the Markov properties imply that every d-separation in a DAG $\mathcal{G}$ induces a conditional independence in the corresponding distribution p. If, in addition, the converse is true and every conditional independence under p corresponds to a d-separation in the DAG, then we have the probabilistic version of *faithfulness*. If the DAG is supplemented with an interpretation in terms of atomic intervention as in Definition 15.3.2, probabilistic faithfulness implies *causal faithfulness*. Hence, every directed edge corresponds to an intervention effect and every absence of a directed edge to the absence of a direct causal effect. Moreover, under causal faithfulness it follows that if, and only if, there is a directed path from X_j to X_k in $\mathcal{G}$ then some manipulation of the former will affect the distribution of the latter.

Faithfulness is often invoked when we want to empirically construct or check parts of the causal model. *Causal search* (or *causal discovery*) refers to methods that reconstruct the whole causal DAG, up to Markov equivalence, based on observational data [67]. Chapter 18 in this book deals with these methods in more detail. We return to faithfulness in Section 15.5.1.

Structural Equation Models

Causal DAGs are often combined with the potential outcomes framework. One way of doing this relies on (recursive) structural equation models (SEMs) [26]. To keep it general we only consider non-parametric ones (NPSEMs) where the shape of the functional relations is left unspecified. These models assume a set of variables to be functionally related at an individual level such that these functional relations are invariant to how the input comes about. Consider $\mathbf{X} = (X_1, \ldots, X_K)$ and a DAG $\mathcal{G}$. A structural equation model for $\mathbf{X}$ on $\mathcal{G}$ assumes for each node $k \in V$ that X_k is a function of its graphical parents and possibly a random variable ϵ_k [47]

$$X_k := f_k(\mathbf{X}_{\mathrm{pa}(k)}, \epsilon_k),$$

where in the simplest case all ϵ_k are assumed mutually independent. A violation of this independence of assumption can graphically be depicted by corresponding bi-directed edges resulting in a so-called *semi-Markovian* graph [49]. We write ':=' to remind the reader that the above should be regarded as an asymmetric *assignment* of the value of f_k to X_k [47]. The joint distribution of $(\epsilon_1, \ldots, \epsilon_K)$ together with the above functional relations induces a distribution on $(X_1, \ldots, X_K)$.

The *structural* nature of such a system means that potential outcomes can be constructed as follows. Consider an intervention forcing X_j to have the value $\widetilde{x}_j$. Under an NPSEM such an intervention corresponds to replacing the function f_j with $\mathbf{1}\{x_j = \widetilde{x}_j\}$, and x_j is replaced by $\widetilde{x}_j$ in the argument of f_k whenever $j \in \mathrm{pa}(k)$. The latter yields equations for the potential outcomes $X_k(\widetilde{x}_j)$. However, instead of *replacing* equations, NPSEMs allow to simply *add* equations for each intervention. The result is a system of equations that simultaneously describes what would happen if X_j was fixed at value $\widetilde{x}_j$ as well as at value

$\widetilde{x}'_j$ etc. Hence, the joint distribution of $(\epsilon_1, \ldots, \epsilon_K)$ also induces a joint distribution on all potential outcomes $\{X_k(\widetilde{x}_j); k \in V, \widetilde{x}_j \in \mathcal{X}_j\}$.

Note that the above combination of DAGs with NPSEMs agrees with Definition 15.3.2 in the sense that with the above construction the joint distribution $p(\mathbf{X}_B(\widetilde{\mathbf{x}}_A))$ will satisfy the definition of $p(\mathbf{x}_B; \mathrm{do}(\widetilde{\mathbf{x}}_A))$. However, a causal DAG as in Definition 15.3.2 does not encompass joint distributions under interventional settings of a variable simultaneously to different values.

Example: Consider the DAG in Figure 15.4(b) to which we return later in the context of bias amplification. Linked with a NPSEM it induces variables C, U, X, Y as follows:

$$C := \epsilon_C, \quad U := \epsilon_U, \quad X := f_X(C, U, \epsilon_X), \quad Y := f_Y(X, U, \epsilon_Y).$$

A set of potential outcomes is defined for any fixed values $\widetilde{c} \in \mathcal{C}, \widetilde{x} \in \mathcal{X}$ as

$$X(\widetilde{c}) := f_X(\widetilde{c}, U, \epsilon_X), \quad Y(\widetilde{x}) := f_Y(\widetilde{x}, U, \epsilon_Y),$$

yielding $4+|\mathcal{C}|+|\mathcal{X}|$ equations / variables in our system if C and X are discrete, and infinitely many if they are continuous. Note that we could extend this to interventions in U yielding even more potential outcomes. Any joint distribution of $(\epsilon_C, \epsilon_U, \epsilon_X, \epsilon_Y)$ induces a joint distribution on all of these, that is on $(C, U, X, Y, \{X(\widetilde{c}) : \widetilde{c} \in \mathcal{C}\}, \{Y(\widetilde{x}) : \widetilde{x} \in \mathcal{X}\})$, where we use that $Y(\widetilde{x}, \widetilde{c}) = Y(\widetilde{x})$. Within this set-up, certain counterfactual quantities are therefore well-defined, such as for instance (for binary C and X) $E(Y(1) - Y(0)|X(1) > X(0))$. In the context of an RCT with imperfect compliance, let C be the randomized treatment and X the actual treatment taken. Then, $E(Y(1) - Y(0)|X(1) > X(0))$ can be regarded as the effect of treatment on the group of people who 'comply' with their assignment as these are characterized exactly by $X(1) > X(0)$. As we can never observe $X(1), X(0)$ together, this is a latent subgroup of the population. Moreover, a common assumption in this context is the one of monotonicity which demands that f_X is such that $X(1) \geq X(0)$ always. Such counterfactual assumptions cannot be read off the graph and are an additional specification of the NPSEM.

The above gives an idea of how extensive a latent structure is imposed by NPSEMs. A less restrictive approach for a graph-based construction of potential outcomes is discussed in Robins and Richardson [55].

15.3.3 Comparison

The differences between the above approaches are subtle and lie partly in the different graphical displays and partly in the different semantics. Intervention DAGs use the additional intervention nodes to supplement the graph with causal meaning, while causal DAGs modify the meaning of edges and separations via the causal Markov properties. With the above definitions, intervention DAGs are the more general class of models. Causal DAGs are a special case where there is an (invisible) intervention node into every variable and all interventions are atomic ones. The NPSEM interpretation of causal DAGs further imposes a counterfactual structure as discussed.

Causal DAGs can and have been generalized to cases where the interventions are other than atomic, for example random, and where not all nodes need to be manipulable. For instance, Definition 15.3.2 can be relaxed by demanding in (ii) that the property holds *locally*, that is only for a *specific* set $A \subset V$ [40, 49]. Such a specific set could for instance be the set of all treatment-like variables that can be manipulated in practice. However, such a relaxation is rarely used in the literature, even though many results that have first been

derived for causal DAGs can be shown to hold more generally for locally causal DAGs. The latter is more explicit in intervention DAGs due to the corresponding intervention nodes. Arguably, in practice, it is often not plausible nor required to assume that all variables are manipulable and we find that using intervention DAGs is a helpful tool to focus on the essential assumptions. For instance, in the next section we give two results, the back-door and the front-door criterion. The former was proposed by Pearl [47] using causal DAGs, but a close look at his proof shows that it uses a simpler intervention DAG as also shown by Dawid [11]. The same is true for the front-door criterion as we demonstrate in Section 15.4.2.

15.4 Graphical Rules for the Identification of Causal Effect

One of the most prominent uses of DAGs in causal inference is to help decide whether and how the available data identifies a desired causal target of inference under the assumed causal model. Chapter 16 in this book deals with this topic in more detail. Here we consider some classic fundamental results. Let us first clarify what is meant by 'identifiability' [56, 67, 49].

Definition 15.4.1 (Identifiability). *Consider $\mathbf{X} = \mathbf{X}_V$ and let (O, U) be a partition of V. Let $\mathcal{P}$ be a class of distributions for $\mathbf{X}$, including the relevant observational and interventional distributions, for instance, but not necessarily, as induced by a causal DAG. Finally, let $A \subset O$ and $D \subset O \backslash A$.*
We say that the causal effect of $\mathbf{X}_A$ on $\mathbf{X}_D$ is identified by $\mathbf{X}_O$ under $\mathcal{P}$ if for any two distributions $p', p'' \in \mathcal{P}$

$$p'(\mathbf{x}_O) = p''(\mathbf{x}_O) \quad \Rightarrow \quad p'(\mathbf{x}_D; \mathrm{do}(\widetilde{\mathbf{x}}_A)) = p''(\mathbf{x}_D; \mathrm{do}(\widetilde{\mathbf{x}}_A)) \quad \forall \widetilde{\mathbf{x}}_A \in \mathcal{X}_A. \tag{15.3}$$

More generally consider a situation where we are interested in comparing a set of regimes $s \in \widetilde{\mathcal{S}}$ based on data from an observational regime $\sigma = \emptyset$. We say that the consequences of regimes $\widetilde{\mathcal{S}}$ for $\mathbf{X}_D$ are identified if for any two distributions $p', p'' \in \mathcal{P}$

$$p'(\mathbf{x}_O; \sigma = \emptyset) = p''(\mathbf{x}_O; \sigma = \emptyset) \quad \Rightarrow \quad p'(\mathbf{x}_D; \sigma = s) = p''(\mathbf{x}_D; \sigma = s) \quad \forall s \in \widetilde{\mathcal{S}}. \tag{15.4}$$

The above says that the effect of atomic interventions, or the consequence of a regime s, is not a function of the unobserved part $\mathbf{X}_U$. Once $\mathbf{X}_O$ is observed, $p(\mathbf{x}_D; \mathrm{do}(\widetilde{\mathbf{x}}_A))$ can uniquely be determined. Typically, identifiability is demonstrated by finding an expression for $p(\mathbf{x}_D; \mathrm{do}(\widetilde{\mathbf{x}}_A))$ (or for $p(\mathbf{x}_D; \sigma = s)$) in terms of the observational $p(\mathbf{x}_O; \sigma = \emptyset)$. Note that identification as defined above concerns the whole interventional distribution $p(\mathbf{x}_D; \mathrm{do}(\widetilde{\mathbf{x}}_A))$ within the whole model class $\mathcal{P}$. In practice, interest may be restricted to certain parameters such as the ACE or causal odds ratio, possibly within a parametric subclass such as linear or logistic models. There are situations where specific causal parameters can be identified without the whole intervention distribution being identified, for example odds ratios in a case-control study [19]. However, we do not go into these details here. Moreover, we ignore any issues due to finite sample size. In Section 15.5.2 we briefly address situations where identification can be achieved for almost the whole model class $\mathcal{P}$ except a lower dimensional subset.

Positivity

A general requirement for identifiability is that there is an empirical basis for estimating the consequences of the contemplated interventions. This means that combinations of values

possible under the interventional regimes must also be possible under the observational regime. A violation could be if we wanted to compare 'treatment' with 'no treatment' in a patient group where some patients are so ill that they are never left untreated in practice. The exact formulation of this assumption depends on the context and the causal target of inference, but usually requires a positive probability for the relevant combinations under the observational regime — hence known as *positivity*. It will generally be assumed to hold in the following.

15.4.1 The Back-Door Theorem

A first identification result based on causal DAGs is known as the 'Back-Door Theorem' [46, 47]. It is closely related to the assumption of 'no unobserved confounding' as addressed in Section 15.5.1. The name stems from the following terminology: a *back-door path* from j to k in a DAG $\mathcal{G}$ is a path that does not have an edge emanating from j. In other words, it is of the shape $j \leftarrow \cdots k$, where $j, k \in V, j \neq k$.

Theorem 15.4.2 (Back-Door Theorem)**.** *Let $\mathcal{G}$ be a causal DAG for $\mathbf{X}$. The causal effect of X_j on X_k is identified by $\mathbf{X}_{\{j,k\}\cup C}$, $C \subset V \backslash \{j, k\}$ in the sense of (15.3) if*
(i) $C \cap \mathrm{de}(j) = \emptyset$,
(ii) C blocks every back-door path from j to k in $\mathcal{G}$.
The intervention distribution is then identified as

$$p(x_k; \mathrm{do}(\widetilde{x}_j)) = \int p(x_k | \widetilde{x}_j, \mathbf{x}_C) p(\mathbf{x}_C) \, \mathrm{d}\, \mathbf{x}_C. \tag{15.5}$$

Equation (15.5) can be recognized as the adjustment or standardization formula, which makes many appearances in the literature with different motivations, see [9, 56, 30, 28, 38]. Note that the type of positivity required here demands that for all $\widetilde{x}_j \in \mathcal{X}_j, \mathbf{x}_C \in \mathcal{X}_C$, we have $p(\widetilde{x}_j, \mathbf{x}_C) > 0$.

The Back-Door Theorem can also be derived from Theorem 7.1 of Spirtes et al. [67]. Using intervention DAGs, Lauritzen [40] and Dawid [11] provide slightly different versions and more general proofs of the above theorem than Pearl [49]. These alternatives demonstrate that the same criterion can be applied without assuming a causal DAG as long as the model is valid (the assumed invariances hold) with regard to an intervention in X_j, or when a corresponding intervention DAG is used for representing an atomic intervention in X_j.

Example ctd. Returning to the examples in Figure 15.1 (omit the node σ to read these as causal DAGs), we find that in (a) C blocks the only back-door path, while in (b) it needs to be supplemented by Z in order to block all three back-door paths. As alluded to earlier, this is an example where it is actually not necessary to decide the causal direction between Z and C, or to assume a causal relation between them at all, for inference to be valid. In fact, Figure 15.1(a) expresses exactly the assumptions of Dawid's [11] version of the Back-Door Theorem, where the interventional regimes target X, and C is characterized by the invariances in (a).

Use of the Back-Door Theorem in Practice

A way to use Theorem 15.4.2 is to construct an assumed causal DAG $\mathcal{G}$, for example based on subject matter background knowledge. *To be plausible this typically involves unobserved quantities.* With the assumed causal DAG, it can be checked whether a set of *measured* covariates $\mathbf{X}_C$ can be found such that C satisfies the back-door criterion relative to (j, k) in

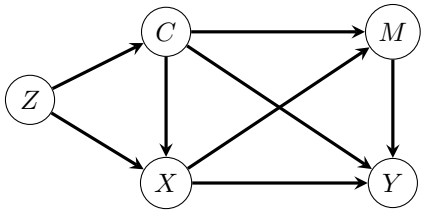

FIGURE 15.2: Assumed causal DAG for end stage renal disease example [68]. Here, X = 'smoker' (at entry in study), Y = 'ESRD', Z = 'early smoker' (unobserved). Further C are observed pre-exposure covariates such as age, sex, ethnicity, prior diseases, and M are post-exposure covariates such as the urinary albumin-to-creatinine ratio, for example.

this DAG. In that case, no further measurements are required to identify any causal effect that is a function of $p(x_k; \mathrm{do}(\widetilde{x}_j))$.

Example: Staplin et al. [68] aim to investigate the effect of the binary indicator X = 'smoker at entry into the study' on the binary indicator Y = 'End Stage Renal Disease' (ESRD), where a number of covariates have been observed, but smoking status in early adulthood is unobserved. First we note that condition (i) $C \cap \mathrm{de}(X) = \emptyset$ of the back-door criterion implies that C consists only of *pre-exposure covariates*. This means that they must be known to not be affected by an intervention in the exposure, for instance because they are prior in time. Condition (i) mirrors the recommendation that for estimating a total causal effect we should not adjust for post-exposure covariates [7]. Without referring to a graph this requirement can be stated as $C \perp\!\!\!\perp \sigma$ or $p(c; \mathrm{do}(X = \widetilde{x})) = p(c)$.[5] Hence, for the elicitation of the assumed causal DAG, subject matter knowledge should be used to determine, first of all, which covariates are pre-exposure. Graphically, these cannot be descendants of the exposure node. In the example, early smoking status, age, sex, ethnicity, and prior diseases cannot possibly be affected by current smoking and are hence considered pre-exposure covariates. In contrast, covariates such as the urinary albumin-to-creatinine ratio and renal status could be affected by current smoking status and are therefore post-exposure. These and further subject matter assumptions are depicted in Figure 15.2. The authors argue that smoking status in early adulthood carries no further information for ESRD once we account for current smoking status and prior diseases, hence there is no edge from Z to Y in Figure 15.2. The assumed causal DAG can now be queried as to whether the measured pre-exposure covariates C satisfy (ii) so that the unmeasured earlier smoking status Z can safely be ignored. This is indeed the case with C including covariates age, sex, ethnicity as well as information on prior diseases [68]. The example illustrates that subject matter knowledge combined with the back-door criterion sometimes leads to the selection of a subset of the covariates as it turns out that not all of them are required for valid causal inference.

Using a DAG in this way strengthens any analysis as it makes explicit and transparent the key assumptions. It can also help to identify weaknesses, for example important unmeasured covariates, and can thus potentially lead to improvements of design and data-collection for future studies.

An apparent limitation of using causal DAGs in the above way is that researchers often

[5] Section 15.4.2 gives a different graphical criterion showing how post-exposure covariates can be used. Section 15.5.1 gives an example for so-called M-bias demonstrating that adjusting for pre-exposure covariates does not protect from introducing bias.

find it difficult to specify the whole DAG, including all pairwise relations between any two variables, observed as well as relevant unobserved variables. For instance, it may not be clear what the causal direction between alcohol consumption and smoking status are as these are processes over time and both reflect a general life-style. In such cases it may be helpful to elaborate the DAG, for example to include 'life-style choice' as an underlying unobservable quantity that encompasses several issues.

Remarks and Generalizations

Theorem 15.4.2 provides a sufficient (graphical) criterion for the identification of $p(x_k; \mathrm{do}(\widetilde{x}_j))$; necessary conditions are given for instance in [64] and Chapter 16 in the present part of this book. However, these are necessary only in the context of causal DAGs. As has been discussed earlier, we may not want to assume causal validity with regard to *all* nodes, but only locally, or we may want to work with intervention graphs. In these cases it is unclear whether necessary conditions can be given.

Perković et al. [52] provide a complete investigation of graphical criteria for identification specifically through the adjustment formula (15.5). They generalize the existing results in a number of ways, for instance showing that sets C exist that do not satisfy (i) but for which (15.5) is still valid. Further, they consider larger classes of graphs than DAGs so as to cover latent variables and relevant Markov equivalence classes.

Sequential Treatments. A further generalization concerns multiple, possibly sequential, interventions. In fact, the Back-Door Theorem can immediately be extended to sets of intervention nodes [46, 47, 52]. Here we consider the additional challenge when variables that are descendants of earlier intervention nodes are needed to identify later interventions. To give a brief idea, consider a time-ordered sequence $(\mathbf{X}_{C_0}, X_i, \mathbf{X}_{C_1}, X_j, X_k)$ and assume we want to know the effect of a joint intervention in (X_i, X_j) on X_k, denoted by $p(x_k; \mathrm{do}(\widetilde{x}_i, \widetilde{x}_j))$. Such situations occur for example in the treatment of chronically ill patients. Let the causal DAG be such that $C_0 \cap \mathrm{de}(i,j) = \emptyset$ and $C_1 \cap \mathrm{de}(j) = \emptyset$ but $C_1 \cap \mathrm{de}(i) \neq \emptyset$ meaning that $\mathbf{X}_{C_1}$ can be affected by interventions in X_i but not X_j. In other words, $\mathbf{X}_{C_1}$ are post-X_i but pre-X_j covariates. It is not obvious how to apply Theorem 15.4.2 when C_1 is needed to block some back-door paths from X_j to X_k but contains at the same time descendants of X_i. Extending the back-door criterion [51, 14], causal DAGs or influence diagrams can again be used to characterize when $(\mathbf{X}_{C_0}, \mathbf{X}_{C_1})$ identify the desired causal effect such that

$$p(x_k; \mathrm{do}(\widetilde{x}_i, \widetilde{x}_j)) = \int p(x_k | \widetilde{x}_i, \widetilde{x}_j, \mathbf{x}_{C_0}, \mathbf{x}_{C_1}) p(\mathbf{x}_{C_1} | \widetilde{x}_i, \mathbf{x}_{C_0}) p(\mathbf{x}_{C_0}) \, \mathrm{d}\, \mathbf{x}_{C_0} \mathbf{x}_{C_1}. \tag{15.6}$$

The above has been termed 'g-formula' by Robins [56] and is a generalization of the adjustment formula (15.5). It is straightforward to further generalize the above such that, for example, $\widetilde{x}_i$ is a function of $\mathbf{x}_{C_0}$ and $\widetilde{x}_j$ a function of $\mathbf{x}_{C_1}$, for instance to represent the case where treatment decisions depend on previous measurements taken on the patient.

15.4.2 The Front-Door Theorem

As mentioned above, the Back-Door Theorem provides a sufficient criterion for the identification of a causal effect. In this section we consider a second sufficient graphical criterion, the Front-Door Theorem [47]. The complete identification algorithm of Shpitser and Pearl [64] can be regarded as combining and generalizing these two criteria in the context of causal DAGs.

We use this opportunity to re-state the theorem not in its original version, but in the

framework of intervention DAGs. This illustrates the use of intervention graphs, and demonstrates that it is not necessary to assume a fully causal DAG to obtain identifiability. The following theorem therefore generalizes the original result. Moreover, we hope that this helps the reader to become more familiar with the intervention DAG framework.

Theorem 15.4.3 (Front-Door Theorem). *Let $\mathcal{G}^\sigma = (V \cup \sigma, E^\sigma\})$ be an intervention DAG for $\mathbf{X}$ under $p(\mathbf{x}; s)$. Let $s = \emptyset$ denote the observational regime and $s = \widetilde{x}_j \in \mathcal{X}_j$ denote an atomic intervention fixing X_j at x_j. Let, further, $C \subset V \backslash \{j, k\}$ and $U \subset V \backslash (\{j, k\} \cup C)$. The consequences of regimes $s = \widetilde{x}_j \in \mathcal{X}_j$ are identified by $\mathbf{X}_{\{j,k\} \cup C}$ in the sense of (15.4) if*
(i) $\mathbf{X}_U \perp\!\!\!\perp \sigma$, and
(ii) $\mathbf{X}_C \perp\!\!\!\perp (\sigma, \mathbf{X}_U) \,|\, X_j$, and
(iii) $X_k \perp\!\!\!\perp (\sigma, X_j) \,|\, (\mathbf{X}_C, \mathbf{X}_U)$.
The intervention distribution is then identified as

$$p(x_k; \sigma = \widetilde{x}_j) = \int p(\mathbf{x}_C | \widetilde{x}_j; \sigma = \emptyset) \int p(x_k | x_j, \mathbf{x}_C; \sigma = \emptyset) p(x_j; \sigma = \emptyset) \, \mathrm{d}\, x_j \, \mathrm{d}\, \mathbf{x}_C. \quad (15.7)$$

Figure 15.3 shows the intervention graph depicting the invariance assumptions of Theorem 15.4.3.

Proof of Theorem 15.4.3: Using all of conditions (i, ii, iii) and that $p(x_j | \mathbf{x}_U; \sigma = \widetilde{x}_j) = \mathbf{1}\{x_j = \widetilde{x}_j\}$ by definition, we have

$$\begin{aligned} p(x_k; \sigma = \widetilde{x}_j) &= \int p(x_k | x_j, \mathbf{x}_C, \mathbf{x}_U; \sigma = \widetilde{x}_j) p(\mathbf{x}_C | x_j, \mathbf{x}_U; \sigma = \widetilde{x}_j) \\ &\quad \times p(x_j | \mathbf{x}_U; \sigma = \widetilde{x}_j) p(\mathbf{x}_U; \sigma = \widetilde{x}_j) \, \mathrm{d}(x_j, \mathbf{x}_C, \mathbf{x}_U) \\ &= \int p(\mathbf{x}_C | \widetilde{x}_j; \sigma = \emptyset) \\ &\quad \times p(x_k | \mathbf{x}_C, \mathbf{x}_U; \sigma = \emptyset) p(\mathbf{x}_U; \sigma = \emptyset) \, \mathrm{d}\, \mathbf{x}_U \, \mathrm{d}\, \mathbf{x}_C. \end{aligned}$$

The claim now follows upon verifying that with the assumed conditional independencies, under the idle regime,

$$\int p(x_k | \mathbf{x}_C, \mathbf{x}_U; \sigma = \emptyset) p(\mathbf{x}_U; \sigma = \emptyset) \, \mathrm{d}\, \mathbf{x}_U = \int p(x_k | x_j, \mathbf{x}_C; \sigma = \emptyset) p(x_j; \sigma = \emptyset) \, \mathrm{d}\, x_j.$$

This last term corresponds to the back-door formula (15.5) for the 'effect' of $\mathbf{X}_C$ on X_k with X_j blocking the only back-door path. In this sense, the front-door formula (15.7) can be interpreted as exploiting the Markov structure so as to combine the effect of X_j on $\mathbf{X}_C$ and the 'effect' of $\mathbf{X}_C$ on X_k. However, as can be seen from the assumptions of the above theorem and its proof, no intervention in $\mathbf{X}_C$ is assumed. The result follows as long as the conditional independencies and invariances of assumptions (i)-(iii) regarding intervention in X_j can be justified.

The Front-Door Theorem appears to be used very little in practice. Its assumptions require that the causal effect of X_j on X_k is known to be 'fully mediated' by $\mathbf{X}_C$ while at the same time this $\mathbf{X}_C$ is known to be conditionally independent of the unobserved $\mathbf{X}_U$ given X_j. With 'fully mediated' we refer to the assumptions of the theorem implying that $p(x_k | \mathbf{x}_C; \sigma = \widetilde{x}_j)$ is in fact not a function of $\widetilde{x}_j$. Pearl [47] gives the following example: let X_j be smoking intensity, X_C the amount of tar deposit in the lungs and X_k an indicator

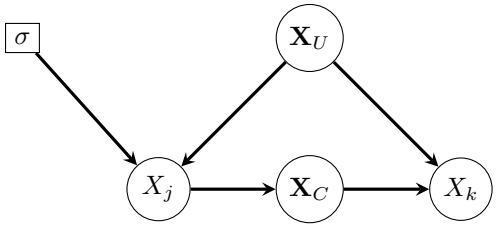

FIGURE 15.3: Intervention graph illustrating the Front-Door Theorem.

for lung cancer. Note that while it may be feasible to intervene in and modify the smoking intensity, it seems much less practical to intervene in the tar deposit without changing the smoking intensity. Hence an approach only considering an intervention in X_j but not X_C seems appropriate. Further, it appears plausible that while smoking itself is related to many observed and unobserved social and life-style factors under the observational regime, the tar deposit is independent of these given smoking intensity. In turn it is plausible that once we account for the actual tar deposit in the lungs smoking intensity does not further contribute to developing lung cancer. Hence the key assumptions are plausible in this situation but the difficulty is to obtain measurements on X_C.

15.5 Graphical Characterization of Sources of Bias

The main threats to the validity of causal inference are threefold: (i) confounding, (ii) measurement error, and (iii) selection bias. Of these, confounding is the one most specific to causal inference, while measurement error or selection also impede consistent estimation of associational or other statistical targets of inference, such as estimating the prevalence of a disease in a population. We therefore focus on the use of graphs especially in the context of confounding, followed by a brief overview of the issues surrounding selection.

15.5.1 Confounding

One of the most fundamental assumptions underlying causal inference is that of 'no unmeasured confounding' [57], also known as 'ignorability' [62], 'exchangeability' [29, 36], or 'exogeneity' [37]. Dawid [11] speaks of a 'sufficient set of covariates' or 'unconfounders', which allow identification of causal targets if measured. Note that the assumption typical concerns conditions for the *absence* of a source of bias, namely 'confounding', but does not define confounders themselves. The former appears to be easier to formalize than the latter. Nevertheless, one can sometimes find the recommendation that we need to adjust for 'all confounders'. Vanderweele and Shpitser [71] give a detailed account of a number of issues around the notion of confounders. Graphs help clarifying many aspects of the assumption, eliminating ambiguities and lead to a better understanding. To see this, we first discuss some definitions of the absence of confounding that do not refer to graphs.

Sufficient Adjustment Set

Assume that we are interested in the causal effect of an exposure X on an outcome Y, and consider a set of pre-exposure covariates C. Depending on the chosen framework, we say

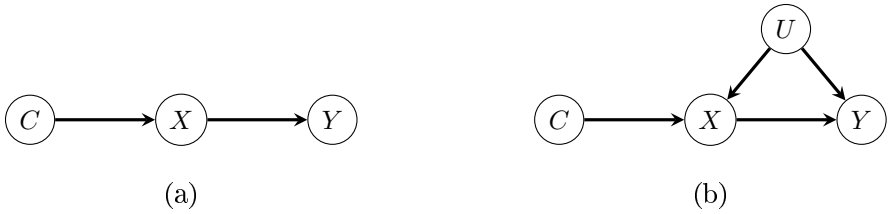

FIGURE 15.4: Causal DAGs: (a) no confounding by C, (b) potential for bias amplification by C.

that there is no confounding of the effect of X on Y given this set of covariates C if

$$\begin{aligned} Y(x) \perp\!\!\!\perp X \,|\, C, \quad \forall x \quad \text{or} \\ p(y|c; \mathrm{do}(\widetilde{x})) = p(y|c, \widetilde{x}), \quad \forall \widetilde{x}, c \text{ or} \\ Y \perp\!\!\!\perp \sigma \,|\, (X, C), \end{aligned} \tag{15.8}$$

where σ takes values in the set of the idle regime and interventions in X. The set C is then *sufficient to adjust for confounding* or a *sufficient adjustment set*. For such a set C it can easily be seen that equation (15.5) is valid with $X = X_j$, $Y = X_k$ and $C = \mathbf{X}_C$. It follows that we can graphically characterize the sufficiency: whether we use a DAG augmented with an intervention node σ or a causal DAG, any set C blocking all back-door paths from X to Y is a sufficient adjustment set. If no proper subset of C satisfies the condition, we call C *minimally sufficient* adjustment set. For there to be 'no unmeasured confounding' at least one such set needs to be measured. If $C = \emptyset$ satisfies the above we say that there is no confounding.

All the above versions of the assumption essentially express that whether X was generated observationally or by an intervention does not further predict Y once C is taken into account. This ensures that inference about the effect of an intervention can indeed be based on data on X, Y, C. The most popular way of adjusting for confounding is based on equation (15.8) and relies on a regression of Y on X and C to estimate $p(y|c; \mathrm{do}(\widetilde{x}))$. The regression coefficient of X receives a causal interpretation as the effect of an intervention $\mathrm{do}(\widetilde{x})$. This approach is known as *regression adjustment*. Strictly speaking, it results in a conditional, or subgroup, causal effect, namely an effect of X for given values of C. This is not necessarily equal to the marginal causal effect as obtained using equation (15.5), where C is integrated out. The distinction is relevant if (i) the effect of X varies for different values of C (effect modification by C), or (ii) the conditional and marginal effects are not identical, a phenomenon known as non-collapsibility (see [7, 28, 38] for discussions).

Examples and Misconceptions

A 'confounder' is sometimes loosely defined as a covariate that predicts both exposure and outcome [71]. This is too imprecise and can lead to overadjustment as seen with the following examples.

Bias Amplification. In Figure 15.4(a), C predicts X and Y but while it does not violate condition (15.8), it is not required because $p(y; \mathrm{do}(\widetilde{x})) = p(y|\widetilde{x})$ meaning there is no confounding in the first place. An extension of this example is given in Figure 15.4(b). Here, C is not a sufficient adjustment set and, moreover, adjusting for C in the presence of an unobserved U may be detrimental: the estimator of the coefficient of X in a linear

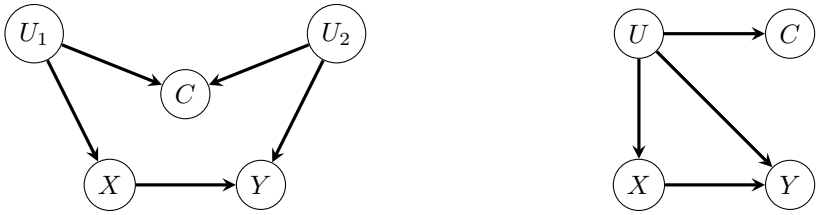

FIGURE 15.5: Causal DAGs: (a) potential for M-bias when adjusting for C, (b) mismeasured confounder.

regression of Y on both X, C will be *more biased* than in a regression of Y on X alone under causal DAG (b). This phenomenon is known as bias amplification [50, 44]. It occurs in linear regressions when covariates are used for adjustment based on strongly predicting exposure X without regard to their relation with the outcome Y as is sometimes recommended in the context of propensity score methods [63].

M-Bias. In Figure 15.5(a), C is a pre-exposure covariate as it is not a descendant of X. Moreover, it may appear to be a confounder as it predicts (or is associated with) both X and Y but, as in Figure 15.4(a), $p(y; \mathrm{do}(\widetilde{x})) = p(y|\widetilde{x})$ so that C is not required to adjust for confounding. In fact, it is again detrimental to adjust for C because as a collider it opens a back-door path so that $p(y|c; \mathrm{do}(\widetilde{x})) \neq p(y|c, \widetilde{x})$. This phenomenon is known as *M-bias* [65] and C is in this case a *bias inducer* [44]. It follows from this example that if a set of pre-exposure covariates is a sufficient adjustment set, a superset will not necessarily be sufficient anymore because additionally conditioning on colliders may open back-door paths. Note that while the exact situation of Figure 15.5(a) may not often be encountered in practice, the one where C additionally has directed edges into X and Y could be quite common. In such a case, a sufficient adjustment set would need to include either U_1 or U_2. This is illustrated with the following application.

Williamson et al. [73] consider a population-based longitudinal cohort study of children. Interest lies in estimating the causal effect of personal smoking (X) on adult asthma (Y), and the question is whether to adjust for childhood asthma (C). Relating this to the DAG in Figure 15.5(a), U_1 is parental smoking, while U_2 stands for underlying atopy. Note that the actual DAG used by these authors to represent subject matter background knowledge is more elaborate with more variables. Fortunately, data on parental smoking U_1 was available so that together with a number of further covariates, including childhood asthma, a plausible measured adjustment set could be found.

Mismeasured Confounder. The causal DAG in Figure 15.5(b) could represent the situation where only an imperfect measurement C of the underlying quantity of interest U is available. While C is associated with both X and Y, it is not a sufficient adjustment set. In fact, if U was available, C would not be relevant. However, with U unobserved, the conditional independence structure as in Figure 15.5(b) characterizes the measurement error as *non-differential*. Then, with additional parametric assumptions (such as lack of effect modification by U, for example), adjusting for C reduces the bias compared to no adjustment. However, as shown by Ogburn and Vanderweele [45], bias reduction is not guaranteed. It requires specific measurement error assumptions that cannot be represented graphically.

FIGURE 15.6: Causal DAGs: (a) both (C_1, C_2) required for adjustment, (b) C_1 or C_2 each sufficient.

Minimal Adjustment Sets. Finally, in Figure 15.6(a) none of C_1, C_2 alone is a sufficient adjustment set, only (C_1, C_2) is, while in Figure 15.6(b) each C_1 or C_2 or (C_1, C_2) are sufficient adjustment sets. Moreover, in (b), each C_1 or C_2 alone are minimally sufficient adjustment sets, demonstrating that minimality does not imply uniqueness.

Selection of Adjustment Sets

We started this section by claiming that the problem of confounding is specific to causal inference. As detailed by Vanderweele and Shpitser [71], no associational definition of confounding is satisfactory. It is evident from the definition of a sufficient adjustment set and from the above examples that the assumption of no unmeasured confounding cannot be tested based on observational data alone as it relates an observational to an interventional distribution. As mentioned in the example after Definition 15.3.1, equation (15.8) could in principle be tested if data from both the observational and interventional regimes were available. We could then compare the conditional distributions to see if they are the same. In practice, the assumption typically needs to be justified based on subject matter knowledge assisted by graphs as demonstrated in the above examples.

Nevertheless, there is clearly a strong motivation to somehow assert empirically that a given set of covariates is sufficient to adjust for confounding, or to empirically *select* a sufficient adjustment set from a possibly very large pool of covariates. This is also an increasingly important topic in the context of high-dimensional data. There are two approaches: (i) First, we can assume that a large set of measured covariates C is a sufficient adjustment set, but we want to determine empirically whether and how this set can be reduced. (ii) We wish to determine empirically whether a large set of measured covariates pre-exposure covariates C contains a sufficient adjustment set.

Reducing a sufficient adjustment set. Assume we are interested in the effect of an intervention in X on Y and let C be a known sufficient adjustment set. Loosely speaking, the reduction of C is based on the idea that variables that do not affect X and (possibly other) variables that do not affect Y can be discarded. In this vein, Robins [58] shows that for disjoint $W_1, W_2 \subset C$ if $Y \perp\!\!\!\perp W_1 \,|\, X, C \backslash W_1$ and $X \perp\!\!\!\perp W_2 \,|\, C \backslash (W_1, W_2)$ then $C \backslash (W_1, W_2)$ is also a sufficient adjustment set. De Luna et al. [15] elaborate this procedure and give two versions of a selection algorithm as well as conditions under which the resulting subsets are minimal. In Figure 15.6(b) the two versions of the algorithm result in reduction either to C_1 or to C_2. Dawid and Guo [31] speak of 'treatment' versus 'response' sufficient reduction of covariates. Vanderweele and Shpitser [71] give a forward selection algorithm which requires causal faithfulness. Certain pitfalls in the empirical selection of covariates in this way can

be illustrated graphically [15]. For instance in the situation of Figure 15.5(a), starting the above algorithms with C will not find that $\emptyset$ is sufficient, while starting the algorithms with (U_1, U_2, C) will correctly conclude that $\emptyset$ is sufficient.

Establishing a sufficient adjustment set. If C is a large set of pre-exposure covariates, but not known to be sufficient for adjustment, Entner et al. [24] invoke an underlying faithful causal DAG model to prove the following. If we can find a set $W \subset C$ as well as $Z \subset (C \backslash W)$ such that $W \not\perp\!\!\!\perp Y|Z$ but $W \perp\!\!\!\perp Y|(Z, X)$, then we can conclude not only that Z is a sufficient adjustment set, but also that an intervention in X has a causal effect on Y. This follows because the two conditions, together with causal faithfulness, imply that there must be directed paths from W to Y blocked by X. Hence there are directed paths from X to Y implying a causal effect, and all back-door paths from X to Y must be blocked by Z. However, when such a W does not exist the question cannot be decided.

15.5.2 Selection Bias

As shown above, confounding can be addressed by conditioning on a suitable set of covariates. Selection bias is the dual problem because it occurs when 'wrongly' conditioning by mistake, design or analysis [34, 7, 19, 5].

When DAGs are used to model and illustrate the problem we are typically alerted to the possibility of selection bias by opening paths due to conditioning on a collider, similar to the phenomenon of M-bias. The problem is therefore also known as *collider-stratification bias* [27]. We refrain from a full treatment here due to space limitation, but give some examples.

Example — Sequential Treatments ctd. Consider the example in Figure 15.7(a), where (X_1, X_2) could be two sequentially administered treatments similar to the situation around equation (15.6). Assume X_1 is randomized but not X_2 so that the latter may be affected by unobserved confounding U. We are interested in the joint effect of (X_1, X_2) on Y, or formally in $p(y; \mathrm{do}(\widetilde{x}_1, \widetilde{x}_2))$. Naively, we might carry out a regression of Y on (X_1, X_2). Unbeknown to us, the true causal structure is as shown. In particular, there is no direct or indirect effect neither of X_1 nor X_2 on Y. Obviously, due to unobserved confounding by U, any effect estimate of X_2 will typically be biased. However, the analysis will also typically find a non-zero effect of X_1 because a joint regression conditions also on X_2. As X_2 is a collider this conditioning opens a path between X_1 and Y. In other words, even without a causal effect of X_1 on Y we have $Y \not\perp\!\!\!\perp X_1|X_2$ as verified via d-separation in Figure 15.7(a). In contrast, a regression of Y on X_1 alone would correctly reveal the absence of any causal effect. Hence, it is the unfortunate choice of a joint regression including X_2 that impedes inference regarding the effect of X_1, a phenomenon pointed out by Robins [56]. In the particular situation of sequential treatments a joint regression is therefore usually not appropriate, and methods such as IPTW or g-computation should be used [59]. Note that an analogous phenomenon applies to the analysis of direct and indirect causal effects. For instance one may attempt to justify conditioning on X_2 with targeting a direct effect of X_1 on Y, but this conditioning introduces selection bias for the same reason as in the sequential treatment example [48, 4].

Sampling Selection

While the above problem occurs due to an unfortunate choice of method for the data analysis, a structurally similar type of problem is due to *sampling selection*. Let S be a binary indicator for being sampled, so that our data is drawn from $p(\mathbf{x}|S = 1)$. It is common to assume random sampling $\mathbf{X} \perp\!\!\!\perp S$ and hence $p(\mathbf{x}|S = 1) = p(\mathbf{x})$. However, many designs

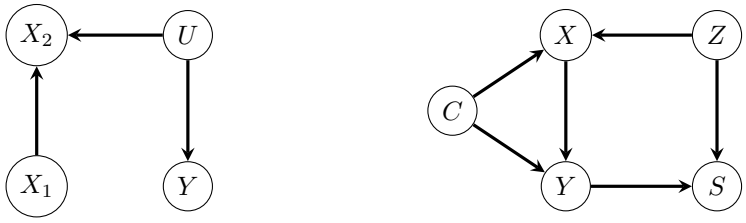

FIGURE 15.7: Causal DAGs with potential for selection bias, (a) sequential treatments; (b) time-to-pregnancy problem.

are based on non-random sampling, for example case-control studies. Hence, we may ask when valid inference about $p(\mathbf{x})$ is possible given data from $p(\mathbf{x}|S = 1)$. For instance, in a case-control study it is well-known that the odds ratio can often still be identified as a measure of association between outcome and exposure. With suitable covariate data, the corresponding conditional odds ratio will also be identified and have a causal interpretation, see overview of Didelez and Evans [18] .

Didelez et al. [19] give graphical conditions for valid causal inference under sampling selection. They consider conditional causal odds ratios as well as the testability of the causal null-hypothesis as their targets of inference. A key issue in this context is that of collapsibility: A quantity being collapsible over S means that inference within the sampled is the same as in the whole population. For instance, the conditional odds ratio between X and Y given C is collapsible over S, $OR_{YX}(C, S) \equiv OR_{YX}(C)$, if, and only if, either $Y \perp\!\!\!\perp S \,|\, (X, C)$ or $X \perp\!\!\!\perp S \,|\, (Y, C)$. Moreover, under these conditions we can also test the null-hypothesis $Y \perp\!\!\!\perp X \,|\, C$ by testing within the sampled, that is conditional on $S = 1$. As these are conditional independence properties, they can easily be checked via separations in a graph. Generalizations of the conditions take more available information into account and hence apply in a wider range of situations. We illustrate this with an example, for a detailed treatment see [19, 2, 3].

Example — Sampling Selection: Let X be exposure to a toxic substance and Y time-to-pregnancy (TTP). This is a common measure of fertility and defined as the time, from initiation, that it takes a couple to become pregnant. As discussed by Weinberg et al. [72], there is a fundamental problem with certain retrospective designs: If couples who are trying to become pregnant or who just gave birth are interviewed during a given time-window, then those with long TTP must have started earlier. Hence, within the sampled (that is, given $S = 1$) initiation time and $Y =$ TTP are necessarily associated. This could induce a non-causal (X, Y)-association if exposure to the substance has changed over time due to legislation, say. The problem is illustrated in Figure 15.7(b), where Z is the initiation time and C is a set of covariates. Key assumptions in the DAG are that $Y \perp\!\!\!\perp Z \,|\, (X, C)$ in the whole population, while $Y \not\perp\!\!\!\perp Z \,|\, (X, C, S)$. It is also assumed that sampling does not depend directly on the exposure so that $S \perp\!\!\!\perp X \,|\, (Y, Z, C)$. With these conditional independencies, it follows from results of [19, 2] that by taking Z into account, we can first collapse over S and subsequently over Z. Hence, the conditional odds ratio in the sample corresponds to the population conditional odds ratio, $OR_{YX}(C, Z, S = 1) = OR_{YX}(C)$. The former can be estimated based on the sampled data, while the latter is the actual target of inference. Similarly, testing for a conditional independence between exposure and outcome given Z and C in the sample, $Y \perp\!\!\!\perp X \,|\, (C, Z, S = 1)$, is equivalent to testing for

conditional independence given only C for the whole population, $Y \perp\!\!\!\perp X \mid C$. A causal interpretation of these odds ratios and tests hinges further on C being a sufficient adjustment set.

The results referred to above concern odds ratios and testing of a relevant null-hypothesis, but they do not achieve full identification of intervention distributions. In the context of sampling selection this is indeed usually not possible [3], but some information can still be extracted from the available data. Evans and Didelez [25] consider *generic identification* of the marginal distribution $p(\mathbf{x})$ from data on $p(\mathbf{x}|S = 1)$, meaning that identification is possible almost everywhere except on a lower dimensional subset of the model parameter space. We do not go into details here, but note that graphical criteria can again be given to characterize when generic identification is possible.

15.6 Discussion and Outlook

In this chapter, we have reviewed key causal concepts in the form of distributional invariances under different regimes, effects of atomic interventions or stable mechanisms. We showed how these can be combined with graphs in different ways. The resulting DAGs can be queried with respect to inferential questions, for instance regarding identifiability of causal quantities or possible sources of bias. There are many further examples where a graphical approach adds clarity by making crucial assumptions explicit and guiding the analysis. A large body of work has developed for instance on the topic of *instrumental variables* which allow some causal inference even when there is unobserved confounding but require very specific and somewhat subtle conditional independence assumptions which need to be carefully justified in any given case [35, 20].

We have only briefly alluded to the converse task, that of causal discovery, where the question is how to determine empirically the causal relations with no or only partial subject matter knowledge (see also Chapter 18 in this book). This is relevant to the estimation of intervention effects especially in high-dimensional settings without prior assumptions [42], or when the whole causal structure is unknown and itself of interest. As an example consider the investigation of genetic regulatory or cellular signaling systems where data from some limited experimental conditions are available and researchers want to determine the most promising future experiments [43]. Numerous causal search algorithms with innumerable variations have been put forward [67], and the field is more active than ever. Naturally this endeavor relies on strong assumptions, such as causal faithfulness in addition to specific parametric assumptions. For critical discussions of causal discovery see Robins et al. [60] and Dawid [12]. Some of the recent developments are concerned with integrating different sources of information, such as data sets that have been obtained under different experimental conditions [22, 32, 70]. For example Peters et al. [53] reverse our earlier reasoning in the following way: If data sets are available from different regimes, such as a number of different experimental settings, their method searches for invariances among conditional distributions to partially reconstruct causal structures. In this context, intervention graphs under general regimes may turn out more useful than causal DAGs [17].

Finally let us point out some extensions to other types of graphs and to dynamic systems. Some generalizations of causal interpretations to a wider class of graphs are motivated by the causal search problem. For instance, so-called maximal ancestral graphs (MAGs) allow for the possibility of latent variables [54, 52]. Other generalizations address the fact that

interventions in dynamic systems cannot necessarily be modeled using DAGs due to the continuous nature of time [8, 1]. Also, when the dynamics are in an equilibrium one may want to employ either partly undirected or cyclic graphs. Hence, an alternative is based on chain graphs [41], where the relevant intervention corresponds to holding fixed a process while the system returns to a new equilibrium. Those alternatives based on cyclic graphs address for example non-recursive structural equation models [66], interventions in time-series [23], or the causal relations among events as modeled by a multi-state process [16]. Here, the cyclicity is usually a short cut for expressing feedback, where the present of one variable depends on the past of another and vice versa. Furthermore, certain dynamic aspects of causal reasoning can alternatively be represented and analyzed with chain event graphs [69]. The fundamental causal concepts and uses of graphs presented in this chapter essentially carry over to all these more complex models.

Bibliography

[1] O. O. Aalen, K. Røysland, J. M. Gran, R. Kouyos, and T. Lange. Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. *Statistical Methods in Medical Research*, 25(5):2294–2314, 2016.

[2] E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 100–108. Journal of Machine Learning Research, 2012.

[3] E. Bareinboim and J. Tian. Recovering causal effects from selection bias. In *Proceedings of the 29th National Conference on Artificial Intelligence*, pages 3475–3481. AAAI Press, Menlo Park, CA, 2015.

[4] S. R. Cole and M. A. Hernán. Fallibility in estimating direct effects (with discussion). *International Journal of Epidemiology*, 31:163–165, 2002.

[5] S. R. Cole, R. W. Platt, E. F. Schisterman, H. Chu, D. Westreich, D. Richardson, and C. Poole. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39:417–420, 2010.

[6] P. Constantinou and A. P. Dawid. Extended conditional independence and applications in causal inference. *Submitted*, 2016.

[7] D. R Cox and N. Wermuth. Causality: a statistical view. *International Statistical Review*, 72(3):285–305, 2004.

[8] D. Dash and M. Druzdzel. Caveats for causal reasoning with equilibrium models. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 192–203. Springer, 2001.

[9] J. A. Davis. Extending Rosenberg's technique for standardizing percentage tables. *Social Forces*, 62:679–708, 1984.

[10] A. P. Dawid. Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association*, 95:407–448, 2000.

[11] A. P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–89, 2002.

[12] A. P. Dawid. Beware of the DAG! *NIPS Causality: Objectives and Assessment*, 6:59–86, 2010.

[13] A. P. Dawid. Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and Its Application*, 2(1):273–303, 2015.

[14] A. P. Dawid and V. Didelez. Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Statististical Surveys*, 4:184–231, 2010.

[15] X. De Luna, I. Waernbaum, and T. S. Richardson. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875, 2011.

[16] V. Didelez. Causal reasoning for events in continuous time: A decision–theoretic approach. In *Proceedings of the 31st Annual Conference on Uncertainty in Artifical Intelligence — Causality Workshop*, pages 40–45, 2015.

[17] V. Didelez. Discussion of: 'Causal inference by using invariant prediction: identification and confidence intervals'. *Journal of the Royal Statistical Society, Series B*, 78(5):990–991, 2016.

[18] V. Didelez and . J. Evans. Causal inference from case-control studies. In N. Breslow, O. Borgan, N. Chatterjee, A. Scott, and G. Mitchell, editors, *Handbook of Case-Control Studies*, Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, 2017.

[19] V. Didelez, S. Kreiner, and N. Keiding. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387, 2010.

[20] V. Didelez and N. A. Sheehan. Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330, 2007.

[21] V. Didelez and N. A. Sheehan. Mendelian randomisation: why epidemiology needs a formal language for causality. In F. Russo and J. Williamson, editors, *Causality and Probability in the Sciences*, volume 5 of *Texts in Philosophy*, pages 263–292. College Publications, London, 2007.

[22] F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

[23] M. Eichler and V. Didelez. On granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, 16(1):3–32, 2010.

[24] D. Entner, P. Hoyer, and P. Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 256–264. JMLR W&CP, 2013.

[25] R. J. Evans and V. Didelez. Recovering from selection bias using marginal structure in discrete models. In *Proceedings of the 31st Annual Conference on Uncertainty in Artifical Intelligence — Causality Workshop*, pages 46–55, 2015.

[26] A. Goldberger. Structural equation methods in the social sciences. *Econometrica*, 40(6):979–1001, 1972.

[27] S. Greenland. Quantifying biases in causal models: Classical confounding vs collider–stratification bias. *Epidemiology*, 14(3):300–306, 2003.

[28] S. Greenland and J. Pearl. Adjustments and their consequences — collapsibility analysis using graphical models. *International Statistical Review*, 79(3):401–426, 2011.

[29] S. Greenland and J. M. Robins. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419, 1986.

[30] S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, 1999.

[31] H. Guo and A. P. Dawid. Sufficient covariates and linear propensity analysis. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 281–288. JMLR W&CP, 2010.

[32] A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B*, 77(1):291–318, 2015.

[33] D. M. Hausman and J. Woodward. Independence, invariance and the causal markov condition. *The British Journal for the Philosophy of Science*, 50(4):521–583, 1999.

[34] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

[35] M. A. Hernán and J. M. Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

[36] M. A. Hernán and J. M. Robins. *Causal Inference*. Chapman & Hall/CRC, 2017. Forthcoming.

[37] G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

[38] N. Keiding and D. Clayton. Standardization and control for confounding in observational studies: A historical perspective. *Statistical Science*, 29(4):529–558, 11 2014.

[39] U. B. Kjaerulff and A. L. Madsen. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Springer Publishing Company, Inc., 2010.

[40] S. L. Lauritzen. Causal inference from graphical models. In O. E. Barndorff-Nielsen, D. R. Cox, and C. Kl'uppelberg, editors, *Complex Stochastic Systems*, pages 63–107. CRC Press, London, 2000.

[41] S. L Lauritzen and T. S Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B*, 64(3):321–348, 2002.

[42] M. H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

[43] F. Markowetz and R. Spang. Inferring cellular networks – a review. *BMC Bioinformatics*, 8(6):22–43, 2007.

[44] J. A. Middleton, M. A. Scott, R. Diakow, and J. L. Hill. Bias amplification and bias unmasking. *Political Analysis*, 24(3):307âĂŞ323, 2016.

[45] E. L. Ogburn and T. J. VanderWeele. Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders. *Biometrika*, 100(1):241–248, 2013.

[46] J. Pearl. Aspects of graphical models connected with causality. In *In Proceedings of the 49th Session of the International Statistical Institute*, pages 391–401, 1993.

[47] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[48] J. Pearl. Direct and indirect effects. In *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 411–420. Morgan Kaufmann, 2001.

[49] J. Pearl. *Causality*. Cambridge University Press, second edition, 2009.

[50] J. Pearl. On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 417–424. Morgan Kaufmann, 2010.

[51] J. Pearl and J. Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 444–453, San Francisco, CA, 1995. Morgan Kaufmann.

[52] E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *ArXiv e-prints*, June 2016.

[53] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, 78(5):947–1012, 2016.

[54] T. Richardson and P. Spirtes. Ancestral graph markov models. *Annals of Statistics*, 30(4):962–1030, 2002.

[55] T. S. Richardson and J. M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Working Paper No.128*, 2013. Center for Statistics and the Social Sciences of the University of Washington.

[56] J. M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods — application to control for the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

[57] J. M. Robins. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79:321–334, 1992.

[58] J. M. Robins and S. Greenland. The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*, 123:392–402, 1986.

[59] J. M. Robins, M. A. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[60] J. M. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

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

[62] D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58, 1978.

[63] D. B. Rubin. Should observational studies be designed to allow lack of balance in covariate distributions across treatment group. *Statistics in Medicine*, 28:1420–1423, 2009.

[64] I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 437–444, Arlington, Virginia, 2006. AUAI Press.

[65] A. Sjölander. Propensity scores and M-structures. *Statistics in Medicine*, 28(9):1416–1420, 2009.

[66] P. Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 491–498, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[67] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT press, second edition, 2000.

[68] N. Staplin, W. G. Herrington, P. K. Judge, C. A. Reith, R. Haynes, M. J. Landray, C. Baigent, and J. Emberson. Use of causal diagrams to inform the design and interpretation of observational studies: An example from the study of heart and renal protection (SHARP). *Clinical Journal of the American Society of Nephrology*, 2016.

[69] Peter Thwaites, Jim Q. Smith, and Eva Riccomagno. Causal analysis with chain event graphs. *Artificial Intelligence*, 174(12):889 – 909, 2010.

[70] I. Tsamardinos, S. Triantafillou, and V. Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. *Journal of Machine Learning Research*, 13:1097–1157, 2012.

[71] T. J. VanderWeele and I. Shpitser. On the definition of a confounder. *The Annals of Statistics*, 41(1):196–220, 2013.

[72] C. R. Weinberg, D. D. Baird, and A. S. Rowland. Pitfalls inherent in retrospective time-to-event studies: The example of time to pregnancy. *Statistics in Medicine*, 12(9):867–879, 1993.

[73] E. J. Williamson, Z. Aitken, J. Lawrie, S. C. Dharmage, J. A. Burgess, and A. B. Forbes. Introduction to causal diagrams for confounder selection. *Respirology*, 19(3):303–311, 2014.