Causal Inference and Bayesian Networks

Anthony Almudevar
Rochester, NY

Contents

Chapter 1

Causal Inference - General Principles

1.1 Introduction

Much of statistical inference concerns relationships between variables. In particular, most basic statistical procedures can be viewed as an attempt to discern *association* between two or more quantitative and/or categorical variables (two-sample tests, ANOVA, linear regression, and so on). For example, if two random variables X and Y are *positively correlated*, then they tend to be either both large or both small when observed as a pair (X, Y). If there is no relationship, then they are independent. One example would be $X =$ `height` and $Y =$ `age` when measured from the same individual.

However, the resolution of scientific questions often requires more than association. *Etiology* is the study of *causality*, and is a central concern to most fields of inquiry. The fact that association by itself is not evidence of causality motivates the field of *causal inference*.

Example 1.1 (Causality in Economics) Suppose economic growth in a certain country follows a specific change in economic policy. It would be natural to ask whether or not the growth is *caused* by that change. Of course, without further evidence, we cannot rule out the possibility that this sequence is merely coincidental. Unfortunately, the only way to resolve the question with complete certainty would be to find another country that is identical in every way, with the exception that economic policy was *not* changed in the manner in question. Then, if the subsequent change in economic growth differs between the two countries, we may conclude that the change in policy does have a causal effect. Obviously, this type of experiment is not feasible, and so the field of *causal inference* is concerned with demonstrating this type of causality in the absence of such ideal data. □

Example 1.2 (Causality in Medicine) In medical science, disease is typically associated with many variables. Some of these will be *causes* of the disease and others will be symptoms (that is, observations *caused* by disease). Clearly, it is important to be able to distinguish between the two. A third type of association occurs when a disease is associated with a variable because both are caused by a third variable. Such a variable is called a *confounder* or *lurking variable*.

A well known example is the association between smoking and cancer. Although smoking is widely believed to be a cause of cancer (in the sense that it increases the chances of developing cancer), for some time it was held as a possibility that both smoking and cancer were caused by a confounder, that is, some third factor which increased the chances both of developing cancer and developing a smoking habit. If this were the case, then smoking behavior would not affect the chances of developing cancer, despite this association (see, for example, Sasco *et al.* (2004)). □

Example 1.3 (Microbes and Disease) Suppose a specific microbe is consistently found in individuals with a certain disease. Furthermore, suppose the converse holds, that is, the microbe tends not to be found in the absence of the disease. Based on this evidence, we might claim that the disease and the microbe are associated. But we have not established that the microbe is the *cause* of the disease. The distinction is crucial, since identification of the microbe as a cause would likely lead to effective preventative or curative measures. On the other hand, establishing association does not rule out the possibility that the presence of the microbe is a *consequence* of the disease, and not a cause, in which case its control will likely have no effect on the disease.

Koch's postulates, formulated in 1884 by Robert Koch and Friedrich Loeffler, consist of four criterion held to imply the existence of a causal relation between a microbe and a disease (Evans, 1978). These postulates specify association of the type just described, but also require that the microbe cause disease when introduced into a healthy individual. The latter condition is intended to rule out the possibility that the microbe follows the disease, rather than precedes it (these postulates are not currently taken as definitive, one reason being that a causal relationship between microbe and disease may require interaction with other factors). □

1.2 Association Versus Causality

The difference between *association* and *causality* is readily understood in an intuitive way, especially when considering specific examples, but can be difficult to define precisely as a general principle. We can say that two events A

and B are *associated* if the occurrence of one is predictive of the occurrence of the other. Similarly, random variables X and Y are *associated* if any observation of one can be used to improve any prediction of the other. This relationship is symmetric.

However, the examples of Section 1.1 make clear that the association by itself cannot be taken as evidence of causality.

Towards a Definition of Causality The exact definition and meaning of causality is, of course, a profound question which is not easily resolved. In the context of empirical investigation, however, we can at least develop a notion of causality which:

(a) Is precise enough to resolve competing hypotheses;

(b) Can be estimated or tested using statistical inference.

If we accept this goal, we can identify three forms of evidence of causality which can be inferred using statistical methods.

(a) Causality as efficacy. Two events A and B are associated, in the sense that they usually occur together. For B to occur it is either necessary or sufficient for A to occur. However, the relationship is not symmetric. The occurrence of A does not depend on the occurrence of B.

(b) Causality as unequal potential outcomes. We wish to discern whether or not a given factor X has a causal effect on a specific outcome Y. We have some ideal list of all factors which have a causal effect on Y. We conduct two experiments which yield observations of Y. In these experiments, all factors are held equal except for X, which by design is forced to differ between the two experiments. Therefore, if the two outcomes Y differ, the only possible cause of this would be X, all other possible causes having been eliminated. Such ideal observations are referred to as *potential outcomes* (Imbens and Rubin, 2015).

(c) Causality and conditional independence. Two random variables X, Y are correlated (or associated), but also independent conditional on a third random variable Z (this concept will be formally defined in Section 2.2). This means that any causal relationship between X and Y is *transient*. If X causes Y, it does so by causing Z, which is then the direct cause of Y (Pearl *et al.*, 2009, 2016).

In the remainder of this chapter we will consider by example these three forms of evidence of causality.

1.3 Causal Inference and the Problem of *Potential Outcome*

Suppose we wish to know which of the two therapies T_s (standard) or T_e (experimental) is most effective in lowering blood pressure. To directly observe either therapy's effect we would measure blood pressure Y_1' at time t_1, administer the treatment (say T'), then measure blood pressure Y_2' at some time $t_2 > t_1$. The drop in blood pressure is then $D' = Y_2' - Y_1'$. Of course, we are interested in comparing the blood pressure drop D' for both treatments. Suppose D_s and D_e are the blood pressure drops following treatments T_s and T_e, respectively. The quantity we are interested in is the *treatment effect*

$$\Delta = D_e - D_s. \tag{1.1}$$

If $\Delta < 0$, we conclude that treatment T_e yields a greater drop in blood pressure than treatment T_s.

However, this seemingly simple procedure faces a considerable obstacle. That is, as a practical matter, D_e and D_s cannot both be observed simultaneously. And the only way we may conclude without ambiguity that treatment T_e is superior to T_s is to compare the blood pressure drop D_e to the blood pressure drop D_s *that would have occurred if the subject had been administered treatment* T_s *instead of* T_e (in the previous statement the treatments may be exchanged). This type of statement is referred to as a *counterfactual*, which describes the consequence of an action other than the one actually taken. This is the ideal experiment, because D_s and D_e would be observed when all possible factors which could affect blood pressure would be identical, except for the treatment. Therefore, if there was a difference between D_s and D_e, it could only be because of the difference in treatment.

Of course, such an observation is not possible, but remains an important concept in causal inference. The effect of a treatment hypothetically (but not actually) administered is referred to as a *potential outcome* (under some conventions both the observed and hypothetical outcome are called *potential outcomes*). Note that a potential outcome cannot be observed by administering each treatment to a single subject at different times, since changes in blood pressure can occur naturally over time, and this could not be ruled out as the source of any observed treatment effect Δ.

However, it must be remembered that the purpose of statistics is precisely to estimate quantities that cannot be observed directly, and causal inference can be viewed in this light. What is needed is to ensure that the relevant quantities are carefully defined in the context of a suitable model of causality, and that any data used in the inference possesses whatever properties are required by that model.

Consider, for example, the treatment effect Δ defined in Equation (1.1). If we are to infer that any improvement in blood pressure reduction is caused by a superior treatment, we must regard Δ as being the difference in blood pressure reduction between the two treatments when all other factors are identical. To emphasize the point, we use the term Δ_c to denote the treatment effect that would occur when this condition holds. In other words, Δ_c is more precisely the expected value of the difference between an observed outcome and its associated potential outcome (or, according to other conventions, the difference between two distinct potential outcomes).

Then suppose we are able to collect data from subjects in the following way. Each subject is administered one of treatments T_s or T_e, but not both. Let $X_1, \ldots, X_{n_1}$ be n_1 observations of D_s from n_1 distinct subjects administered treatment T_s, and let $Y_1, \ldots, Y_{n_2}$ be n_2 observations of D_e from n_2 distinct subjects administered treatment T_e. It would seem natural to estimate Δ_c as:

$$\Delta_c \approx \hat{\Delta}_c = \bar{Y} - \bar{X},$$

where $\bar{X}, \bar{Y}$ are the respective sample means of the two samples. Note that Δ_c is a parameter, and is therefore constant, while $\hat{\Delta}_c$ is a random estimator. This means that the important question is whether or not $\hat{\Delta}_c$ is *unbiased*, meaning that

$$E[\hat{\Delta}_c] = \Delta_c. \tag{1.2}$$

This seems to be a reasonable assumption. However, it is worth considering reasons why Equation (1.2) might *not* hold. An important concept in causal inference is *treatment assignment*. We can think of each of the total $n_1 + n_2$ subjects being *assigned* one of the two treatments. In this case, n_1 (or n_2) subjects are assigned treatment T_s (or T_e), resulting in observations $X_1, \ldots, X_{n_1}$ (or $Y_1, \ldots, Y_{n_2}$). This assignment process might be a planned component of the study. On the other hand, the assignment might not be carried out explicitly. It may be the case that the data is collected from previous clinical records, so that the assignment was carried out before the study was planned.

In either case, an important question arises. Is the treatment assignment independent of the potential outcomes? Suppose that, from caution, the experimental outcome T_e is only assigned to healthier individuals. This may have the effect of decreasing the magnitude of D_e (since healthier subjects would tend to have lower blood pressure *prior* to any treatment, so that, on average, any decrease in blood pressure would be smaller for this group). Treatment assignment is therefore *not* independent of potential outcomes. The effect is that the conditions under which each sample was collected differ in a systematic way, so that Δ_c cannot be reliably estimated. In fact, under this assignment rule, if there was in reality no difference between the two treatments, we would expect $\Delta_c = 0$ but would observe $\hat{\Delta}_c \approx E[\hat{\Delta}_c] > 0$.

How then can we construct an unbiased estimate of Δ_c? Recall that in the ideal experimental, when observing D_s and D_e all factors which could influence blood pressure are identical, except for the treatments. In the previous example, the assignment rule ensured that one factor in particular introduced a systematic bias in the outcome of one treatment, but not the other.

Now, suppose for each subject i assigned treatment T_s we can represent the cumulative affect, or bias, of all factors which can influence blood pressure by the quantity δ_i^x. Let δ_j^y be the corresponding quantity for the jth subject assigned treatment T_e. We will then take X_i, Y_j to be idealized measurements of D_s, D_e, respectively, that is, assuming that all conditions are identical except for the treatment. We can then write

$$\hat{\Delta}_c = \left[\bar{Y} - \bar{X}\right] + \left[\bar{\delta}^y - \bar{\delta}^x\right],$$

where $\bar{\delta}^x, \bar{\delta}^y$ are the sample means of the respective treatment biases. Then $\left[\bar{Y} - \bar{X}\right]$ would be an unbiased estimate of Δ_c, but, of course, this would not be observable. On the other hand, $\hat{\Delta}_c$ *is* observable, and if we can ensure that

$$E\left[\bar{\delta}^y - \bar{\delta}^x\right] = 0, \tag{1.3}$$

then $\hat{\Delta}_c$ would be an unbiased estimate of Δ_c (that is, Equation (1.2) would hold).

This means the crucial step is to ensure that Equation (1.3) holds. Its meaning is that while bias introduced by various factors cannot be eliminated, its effect can be balanced, and therefore cancelled, if it can be allocated equally to both treatment groups. There are a number of ways to do this. The *randomized experiment* is an essential method in medical science. This was in use as early as the 19th century, and statisticians generally rely on the mathematical framework introduced in 1923 by Jerzy Neyman (Rubin, 1990). The essential feature of the randomized experiment is that subjects are randomly assigned treatments. In this way, the bias δ_i associated with the ith subject is equally likely to contribute to $\bar{\delta}^x$ or $\bar{\delta}^y$, so that Equation (1.3) will hold.

Another method used to ensure that Equation (1.3) holds is *matching*. Factors which may affect treatment outcomes are first identified (age or gender, for example). Subjects are then grouped in pairs based on similar factor values, with paired subjects randomly assigned different treatments. The biases δ_i^x and δ_j^y of paired subjects would be similar, and, being assigned to different treatments, would on average cancel. This would ensure that Equation (1.3) holds.

1.4 Observational Versus Experimental Data

Experimental Observation of Causality The phrase "A causes B" often implies the existence of some mechanism by which B necessarily occurs when A does, or by which B cannot occur unless A does (the possibility that A interacts with other causes to this effect must ultimately be acknowledged). This is what was referred to in Section 1.2 as *efficacy*.

If we treat "A causes B" as a hypothesis, this might be resolvable using an experimental platform that allows control of the occurrence of A or B. Consider the next example.

Example 1.4 (Chemical Reaction - Version 1) Suppose we observe some system in which two random variables X_1 and X_2 are observed. To fix ideas, suppose the system is a chemical process, with $X_i = 1$ if agent A_i, $i \in \{1, 2\}$ is detected, and $X_i = 0$ otherwise. Then suppose after some number of experimental trials we observe

$$\begin{aligned} P(X_1 = X_2 = 1) &\approx 0.75, \\ P(X_1 = X_2 = 0) &\approx 0.25. \end{aligned} \tag{1.4}$$

We observe statistical variation of the outcome, but also some structure. A reasonable inference would be that the two agents are associated in some way, as though both are part of a common reaction. Furthermore, this reaction occurred in approximately 75% of the trials. Note that these probabilities imply that agents A_1, A_2 are either both present or both absent, since they sum to one.

However, this says nothing about any causal relationship between the two agents. It may be important to know if one agent "causes" the other, in an asymmetric relationship. Put another way, it is possible that the presence of, for example, A_2 depends on the prior presence of A_1, but that the presence of A_1 does not depend on the prior presence of A_2.

The data we have described so far consists of replications from a fixed set of experimental conditions. This is an example of *observational data*. This data can detect association, but cannot resolve a causal hypothesis. Suppose, however, that experimental techniques exist which can either force or suppress the presence of either agent, irrespective of other system variables (which we refer to as *perturbation experiments*). It is helpful to adopt a distinct notation for experimentally determined states. We will write $X_i = +$ if the presence of agent A_i is experimentally forced, and $X_i = -$ if the presence of agent A_i is experimentally suppressed. This would be an example of *experimental data*.

We may then conduct further experimental replications, under the four experimental conditions:

$$\begin{aligned} X_1 &= - \\ X_1 &= + \\ X_2 &= - \\ X_2 &= +. \end{aligned} \tag{1.5}$$

Possibly, the following probabilities are estimated from the resulting experimental data:

$$\begin{aligned} &P(X_2 = 1 \mid X_1 = -) \approx 0, \\ &P(X_2 = 1 \mid X_1 = +) \approx 1, \\ &P(X_1 = 1 \mid X_2 = -) \approx 0.75, \\ &P(X_1 = 1 \mid X_2 = +) \approx 0.75. \end{aligned}$$

What does this tell us? From the estimates given in Equation (1.4) the marginal probability $P(X_1 = 1) \approx 0.75$ was observed. When A_2 is experimentally controlled the marginal probabilities remain

$$P(X_1 = 1 \mid X_2 = -) = P(X_1 = 1 \mid X_2 = +) \approx 0.75,$$

no matter what the experimental controlled value of X_2 is. This means the presence or absence of A_1 was not dependent on the presence or absence of A_2.

On the other hand, whenever the presence of A_1 was experimentally forced, A_2 was also detected (since $P(X_2 = 1 \mid X_1 = +) \approx 1$), and whenever the presence of A_1 was experimentally suppressed A_2 was not detected (since $P(X_2 = 1 \mid X_1 = -) \approx 0$).

We can therefore infer that A_1 causes A_2, a claim which is stronger than the association of A_1 and A_2.

□

Remark 1.1 The explicit distinction between, for example, the expressions $X_1 = 1$ and $X_1 = +$ of Equation (1.5) is the basis for *do-calculus*. See, for example Pearl (1995, 2012) (the term *calculus of intervention* is used in the earlier reference). This constitutes a formal analytical system which supports the type of analysis required by Example 1.4. ///

Example 1.4 makes clear that the form of data has important implications regarding the ability to infer causality.

Definition 1.1 (Observation vs. Experimental Data) Suppose we are able to sample replications of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$. *Observational data* consists of sampled replications of $\mathbf{X}$ from a single joint distribution. *Experimental data* consists of multiple samples of $\mathbf{X}$ from distinct joint distributions, each induced by experimentally constraining the values of one or more variables in $\mathbf{X}$. ///

Can Causality be Inferred with Observational Data? It is an important fact that the causal hypothesis of Example 1.4 cannot be resolved by observational data. We will see why this is the case in our discussion below. So it's worth asking what types of causal hypotheses, if any, can be resolved using observational data. Much of the theory introduced in these notes is motivated by this question.

In fact, the theory turns out to be rich enough that we may consider a somewhat more ambitious question:

How Much Causality ... ? It might be said that what makes a careful study of the theory of causal inference important is the fact that observational data is often able to partially (but not completely) resolve the causal structure of a process. Of course, it might be the case that auxiliary information of some kind can be used to complete the causal model, so even a partial resolution of causality may suffice. Consider a second example.

Example 1.5 (Chemical Reaction - Version 2) We consider an experimental environment similar to that of Example 1.4, but we now have three agents, a, b and c. It is known that they occur in a common reaction. The sequence is unknown, and its resolution would solve an important scientific question. Essentially, we have six hypotheses:

$$a \to b \to c$$
$$a \to c \to b$$
$$b \to a \to c$$
$$b \to c \to a$$
$$c \to a \to b$$
$$c \to b \to a.$$

Suppose we make no use of perturbation experiments, so that only observational data is available.

What Observational Data Can Do In Chapter 2 we will introduce the *Bayesian network model* as a method of representing in a compact and intuitive way the dependencies which exist among a set of random variables

$X_1, \ldots, X_n$. As we will see, using observational data, a Bayesian network model could be used to rule out some, but not all, of these six hypotheses. In particular, this model could identify the middle agent, by establishing that the occurrence or concentration of the first and last agents are *conditionally independent*, given the middle agent (this idea is formally introduced in Definition 2.3).

What Observational Data Can't Do Identifying the middle agent reduces the number of hypotheses from six to two. For example, if we knew that c was the middle agent, we would know that one of the following two hypotheses was correct:

$$a \to c \to b$$
$$b \to c \to a.$$

However, a Bayesian network model estimated using observational data would not be able to distinguish between the remaining two hypotheses. □

Randomized trials and perturbation experiments provide more direct evidence of causal relationships than does observational data. Of course, such data tends to be specialized, and usually requires dedicated effort and expense to acquire. Observational data, on the other hand, is much more widely and conveniently available, which provides the motivation for the development of statistical methods able to resolve causal hypotheses using this form of data. General classes of such methodology include *probabilistic graphical models* (Koller and Friedman, 2009) and *structural equation models* (Kline, 2015). These methods share an emphasis on the careful analysis of multivariate distributional forms and conditional independence; as well as diagrammatic representations of models using graph theory (Pearl, 1995; Greenland *et al.*, 1999)

The first class of such methodology we consider will be the Bayesian network.

Chapter 2

Bayesian Network Models

Bayesian network models are a widely used means of modeling causal relationships using observational data. It does this by constructing a coherent system of conditional independence constraints among any number of random variables. It is described as a type of *graphical model* because the set of conditional independence constraints is explicitly represented by a class of graphs known as *directed acyclic graphs* (DAGs). Thus, a Bayesian network is formally a joint distribution $g(x_1, \ldots, x_n)$ on random variables $X_1, \ldots, X_n$ which satisfies conditional independence constraints imposed by a DAG G.

Much of the theoretical foundation of Bayesian networks was formulated by Judea Pearl (Pearl, 1985, 1986b,a; Pearl *et al.*, 1989; Pearl, 2014). In addition, comprehensive treatments of graphical models can be found in, for example, Lauritzen (1996), Cowell *et al.* (2006) or Koller and Friedman (2009). A more informal discussion which emphasizes applications in the `R` statistical computing environment is offered in Højsgaard *et al.* (2012) or Nagarajan *et al.* (2013). The `R` package `bnlearn` is a remarkably comprehensive collection of functions and utilities supporting Bayesian network modeling, and is a highly recommended resource for the study of this topic (Scutari, 2010).

2.1 Basic Graph Theory

The study of Bayesian networks requires a familiarity with some basic concepts of graph theory. We first motivate the use of graphs by an example.

Example 2.1 Suppose an individual visits a hospital. It is conceivable that this affects the probability that this individual will miss work the following day. We may isolate a relevant set of events or states-of-nature which have

some bearing on the matter, and represent them by the following variables:

$$\begin{aligned} X_1 &= \textit{Visits Hospital} \\ X_2 &= \textit{Exposure to Bacteria} \\ X_3 &= \textit{Immunity to Bacteria} \\ X_4 &= \textit{Acquires Infection} \\ X_5 &= \textit{Resistance to Antibiotics} \\ X_6 &= \textit{Misses Work.} \end{aligned}$$

We leave as an open question the variable types of $X_1, \ldots, X_6$. They may be categorical, binary (`TRUE/FALSE`), quantitative, or some combination of these. For example, X_1 or X_6 could simply be `TRUE` if the individual visits the hospital, or misses work, respectively. However, in a more refined model X_1 might be a *quantitative variable* equal to the *length* of the visit, or X_6 might be a *probability* of missing work. As will be seen, in general graphical modeling admits considerable flexibility on this question.

However, the first step in understanding graphical models and causality is in understanding the forms of dependence between $X_1, \ldots, X_6$. In the present example, this dependence is induced by a natural cause and effect sequence. For example, `Visits Hospital` precedes `Exposure to Bacteria`, since in this model we are interested in exposures which are *caused* by the hospital visit (and which could not have taken place otherwise). Similarly, `Acquires Infection` precedes `Misses Work`, since it would be a *cause* of missing work.

It seems a simple matter, then, to represent our model graphically by assigning a *node* to each variable, and to draw a *directed edge* (or an *arrow*) between nodes with a causal relationship, with the arrow orientation signifying the causal relationship directionally. This is shown in Figure 2.1. □

2.1.1 Mathematical definition of a graph

We now give the mathematical definition of a *graph*, and introduce some terminology to be used throughout these notes.

Definition 2.1 (Definition of a Graph and its Components) Formally, a *graph* is a pair $G = (V, E)$, where V is a set of *nodes* or *vertices*, and E is a set of *edges*, or pairs of nodes. The pairs may be *ordered* (resulting in a directed edge) or *unordered* (resulting in an undirected edge). Nodes are usually *labeled*, in which case they are considered to be distinguishable. A *directed (undirected) graph* contains only directed (undirected) edges. The theory of Bayesian networks is sometimes concerned with graphs that contain

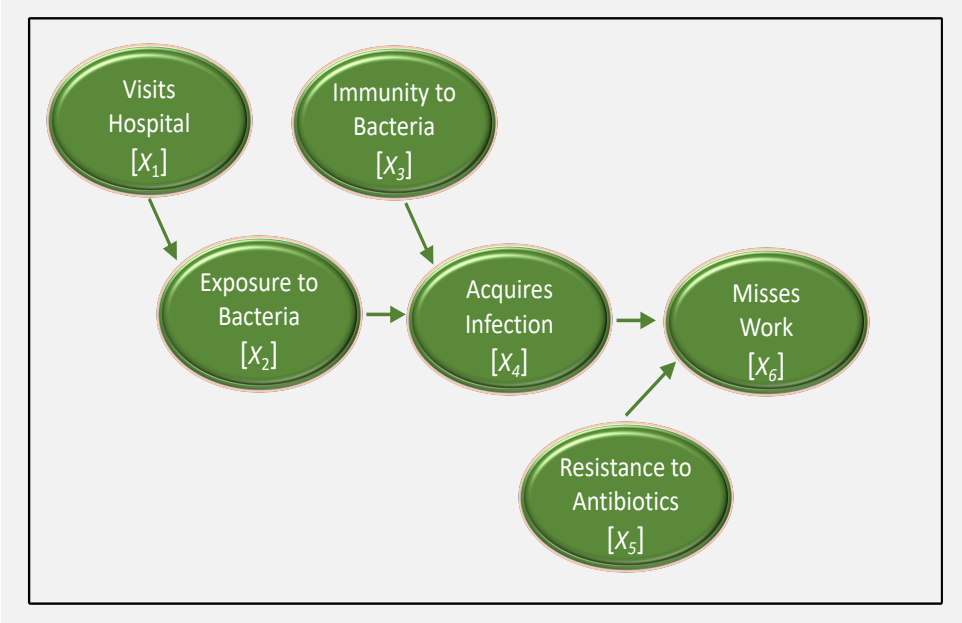

Figure 2.1: Graphical representation of hospital visit (Example 2.1).

both kinds of edges (an example of this is given in Figure 2.11). The graph of Figure 2.1 is a *directed graph*.

We then introduce the following terminology:

- A *subgraph* of $G = (V, E)$ is any graph $G' = (V', E')$ such that V' is a nonempty subset of V, and E' is a subset of E containing only edges which join nodes in V'.
- A directed edge points from a *parent* to a *child*.
- In any graph, a *path* is a sequence of edges which join a sequence of nodes, which, except possibly for the first and last, are distinct. These nodes may be considered to be a component of the path. If a and b are the first and last nodes of the sequence, we say the path *joins* a and b, or is a path *between* a and b.
- In a *directed path* consecutive nodes (left-to-right) form parent/child pairs, and the edges are those which define that relationship. An example from Figure 2.1 is $X_3 \to X_4 \to X_6$. In contrast, $X_4 \to X_6 \leftarrow X_5$ is not a directed path, since X_6 is not a parent of X_5. If a and b are the first and last nodes of a directed path, we say the path is *from* a to b, or *joins* a to b.
- In an *undirected path*, or *arc*, the edges may be in either direction. A *directed path* is also an *arc*. From Figure 2.1, $X_4 \to X_6 \leftarrow X_5$ is an *arc*.

- A *directed path* joins an *ancestor* to a *descendant*.
- A node with no parent is a *founder* or *source*. A node with no children is a *terminal node* or *sink*. The *in-degree* of a node is the number of parents, and the *out-degree* of a node is the number of children. A node is a *founder* if and only if it has in-degree 0. A node is a *terminal node* if and only if it has out-degree 0.
- Let P_j be the set of parents of node j. If node j is a founder, it has no parents, which which case we write $P_j = \{\} = \emptyset$. Similarly, let C_j be the set of children of node j. If node j has no children we may write $C_j = \{\} = \emptyset$. Finally, let D_j be the set of all descendants of node j (this does not include node j itself). If node j has no descendants we may write $D_j = \{\} = \emptyset$. ($\emptyset$ is the conventional symbol for an *empty set*.)
- A path is a *cycle* if the first and last node are the same, with all other nodes appearing at most once. A cycle is either directed or undirected according to the path defining it.
- A *directed acyclic graph (DAG)* is a directed graph that contains no directed cycle (it may contain an undirected cycle). A DAG must contain at least one source and one sink. Figure 2.1 is a DAG.

///

Note that the theory of Bayesian networks is primarily concerned with graphs consisting of nodes labeled with random variables. It will be convenient, therefore, to sometimes refer to the nodes by their random variable labels. Any indices can refer to both a node and to its labeling random variable. For example, random variable X_3 will label a node with index 3. We might then refer to "node X_3", or node $j = 3$. Alternatively, nodes may be labeled with standard variable symbols, such as a, b and c.

Example 2.2 The following is a partial list of the relationships given in Figure 2.1:

- Nodes X_1, X_3, X_5 are *founders*.
- X_5 is a *parent* of X_6; X_6 is a *child* of X_5.
- X_1 is an *ancestor* of X_4; X_6 is a *descendant* X_2.

In addition we have the parent sets

$$P_1 = \{\},\ P_2 = \{X_1\},\ P_3 = \{\},\ P_4 = \{X_2, X_3\},\ P_5 = \{\} \text{ and } P_6 = \{X_4, X_5\},$$

and child sets

$$C_1 = \{X_2\},\ C_2 = \{X_4\},\ C_3 = \{X_4\},\ C_4 = \{X_6\},\ C_5 = \{X_6\} \text{ and } C_6 = \{\}.$$

□

2.1.2 Sequential structure and causality - The directed acyclic graph (DAG)

QUESTION If an arrow denotes causality, as it seems to do, and if in Example 2.1 a visit to a hospital may cause one to miss work, why is there no edge joining node `Visits Hospital` $[X_1]$ to node `Misses Work` $[X_6]$?

ANSWER Put another way, why is node X_6 not a child of node X_1, if a causal relationship clearly exists? This is because a parent-child relationship is not the only means of expressing causality. Note that X_6 is a *descendant* of X_1, which also implies causality. However, the dependence of node X_6 on node X_1 relies on at least one other intervening node. This is known as *transitive causality*.

Distinguishing between transitive and direct causality is a crucial part of causal modeling and inference.

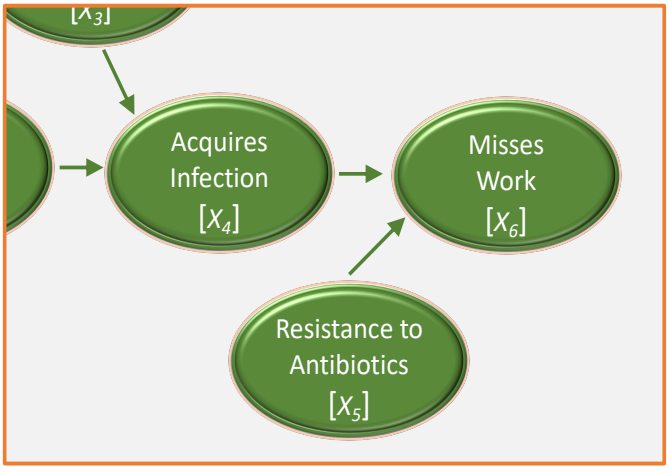

Figure 2.2: Section of Figure 2.1 (parallel sequences).

A DAG represents not just a single sequential process, but possibly several parallel sequential processes, which are only partially synchronized (Figure 2.2). Regarding Example 2.1 and Figure 2.1 we make the following observations:

- We can, at least conceptually, assign a time T_j to each node j, representing the first time X_j is observable. Even without knowing their exact values, they can, to some degree, be ordered.
- In particular, we can claim the following:

$$T_1 < T_2 < T_4 < T_6 \text{ and } T_3 < T_4 \text{ and } T_5 < T_6.$$

- However, the times T_i are only *partially ordered*. Informally this means: (1) we cannot have both $T_i < T_j$ and $T_i > T_j$; and (2) the orderings are *transitive*, so that if $T_i < T_j$ and $T_j < T_k$, we must also have $T_i < T_k$.
- For example, we can say that the statements $T_4 < T_6$ and $T_5 < T_6$ are true, but we cannot say whether $T_4 < T_5$ or $T_5 < T_4$ is true.

THE ROLE PLAYED BY FOUNDERS Note that T_6 is the time at which it can be established that "work is missed". This time can, at least in principle, be precisely identified. On the other hand, the presence, or absence, of a resistance to antibiotics (node X_5) is more of a fixed state-of-nature. What is important to the model is the identity of this state *prior* to T_6. So T_5 can be interpreted as any point of time at or before which the state might be relevant to the outcome. In particular, $T_5 < T_6$. In this case, T_5 cannot be precisely identified, but it is still subject to ordering, which is all that is needed to ensure that this component of the model is positioned coherently. Such nodes, which represent states-of-nature, tend to appear in Bayesian networks as founders.

It is useful to see a graph formed by a pedigree as an example of a DAG (Figure 2.3). The conventional terms *parent*, *child*, *ancestor* and *descendant* used for DAGs conform to their intuitive meanings with respect to a pedigree.

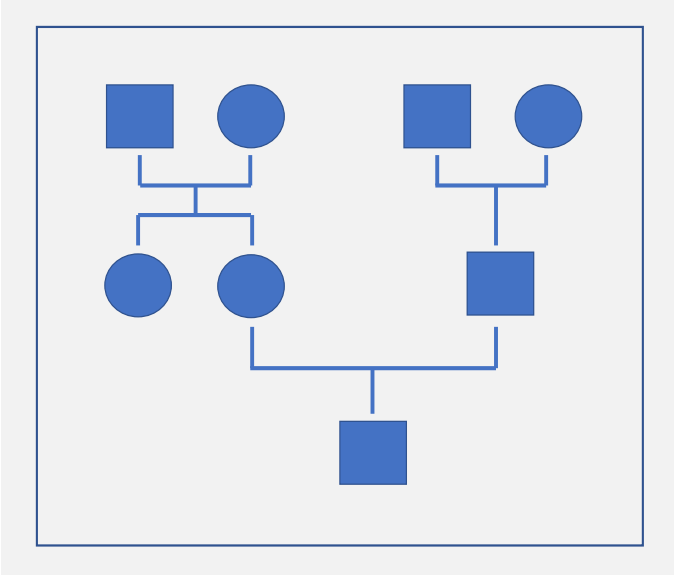

Figure 2.3: A pedigree graph is a DAG.

Definition 2.2 A *topological ordering* of a DAG is an ordering of the nodes with the following property: If node a is a parent of node b, then a precedes

b in the topological ordering. (In some definitions of a *topological ordering* the ordering may be reversed.) ///

Example 2.3 For the DAG of Figure 2.1, the following is a topological ordering:

$$X_1, X_2, X_3, X_4, X_5, X_6.$$

However,

$$X_3, X_5, X_1, X_2, X_4, X_6$$

is also a topological ordering of the same DAG. In general, topological orderings are not unique. □

Example 2.4 Describe the type of DAG for which (a) the topological ordering is unique; (b) all node orderings are topological orderings.

(a) A DAG can consist of a single directed path from a founder to a terminal node (an example is given in Figure 2.6). The topological ordering of such a DAG is unique.

(b) At the other extreme, a DAG defined on nodes V may possess no edges. In this case, all node orderings satisfy the definition of a topological ordering.

□

Example 2.5 For a pedigree, a topological ordering is easily created by ordering the nodes in decreasing order of the age of the individuals represented by the nodes. □

2.2 Conditional Independence

For Bayesian networks, and other causal models based on observational data, causality is a consequence of *conditional independence*. To introduce the idea, we will review the Markov chain model, since it embodies a simplified form of the type of causality captured by a Bayesian network.

2.2.1 The Markov chain model

The probabilistic structure of the Bayesian network can perhaps be made clear by comparison to a *Markov chain* (which, as we will see, is actually a an example of a Bayesian network). Recall that a Markov chain is sequence of random variables $Z_1, Z_2, Z_3, \ldots$ possessing the *memoryless property*:

$$\begin{aligned} P(Z_n = a_n \mid Z_{n-1} = a_{n-1}, Z_{n-2} = a_{n-2}, \ldots, Z_1 = a_1) \\ = P(Z_n = a_n \mid Z_{n-1} = a_{n-1}). \end{aligned} \tag{2.1}$$

The process $Z_1, Z_2, Z_3, \ldots$ is usually understood to unfold in time, so that Z_3 cannot be observed until Z_2 is observed, which cannot be observed until Z_1 is observed. Note that in some definitions of a *Markov chain* Z_i assumes a value from a discrete set, but the distinction is not important here. See Ross (2014) (introductory) or Ross (1996) (more advanced) for excellent introductions to this topic.

PREDICTION PROBLEM Suppose we wanted to predict the value of Z_n, assuming that all previous history $H_{n-1} = (Z_1, \ldots, Z_{n-1})$ is available. In particular, we want to develop some function of H_{n-1}, say S, such that $S(Z_1, \ldots, Z_{n-1}) \approx Z_n$ in some appropriate sense. The prediction will be statistical, and therefore include stochastic error. The best we can do is to make use of the distribution of our target Z_n conditional on all available information, which is in this case H_{n-1}. This distribution is $P(Z_n \mid H_{n-1})$, which is equivalent to the left side of Equation (2.1).

However, because of the memoryless property, expressed mathematically as Equation (2.1), all information in the history H_{n-1} which can be used to predict Z_n is contained in the observation Z_{n-1} alone. So the prediction can be based on the simpler distribution $P(Z_n \mid Z_{n-1}) = P(Z_n \mid H_{n-1})$. This means we would lose nothing by using a simpler predictor of the form $S'(Z_{n-1}) \approx Z_n$ that depends only on Z_{n-1}.

QUESTION Does this mean that Z_n is dependent on Z_{n-1}, but independent of all other observations $Z_{n-2}, Z_{n-3}, \ldots, Z_1$?

ANSWER No. It means that Z_n is *conditionally independent* of observations $Z_{n-2}, Z_{n-3}, \ldots, Z_1$, given Z_{n-1}.

This idea will be made precise in the next section.

2.2.2 Formal definition of conditional independence

We now formally define *conditional independence*.

Definition 2.3 Random events A and B are *unconditionally independent*, or simply *independent*, if

$$P(A \cap B) = P(A)P(B).$$

This condition can be alternatively given by the identity

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = P(A).$$

This may be written $A \perp\!\!\!\perp B$.

Random events A and B are *conditionally independent*, given event C, if

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C).$$

This may be written $(A \perp\!\!\!\perp B) \mid C$.

Random variables X and Y are *unconditionally independent*, or simply *independent*, if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y),$$

where $F_{X,Y}$, F_X, F_Y are the joint and marginal cumulative distribution functions (CDF) of X, Y. This may be written $X \perp\!\!\!\perp Y$.

Random variables X and Y are *conditionally independent*, given Z, if

$$F_{X,Y|Z}(x, y \mid z) = F_{X|Z}(x \mid z)F_{Y|Z}(y \mid z),$$

where $F_{X,Y|Z}$, $F_{X|Z}$, $F_{Y|Z}$ are the joint and marginal cumulative distribution functions (CDF) of X, Y conditional on $Z = z$. This may be written $(X \perp\!\!\!\perp Y) \mid Z$.

Furthermore, the definition extends to groups of random variables. Let $\mathbf{X} = \{X_1, \ldots, X_{n_1}\}$, $\mathbf{Y} = \{Y_1, \ldots, Y_{n_2}\}$ and $\mathbf{Z} = \{Z_1, \ldots, Z_{n_3}\}$ be three groups of random variables, where n_1, n_2, n_3 are three positive integers. Then $\mathbf{X}$ and $\mathbf{Y}$ are *conditionally independent*, given $\mathbf{Z}$, if

$$\begin{aligned} &F_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2} \mid z_1, \ldots, z_{n_3}) \\ &= F_{\mathbf{X}|\mathbf{Z}}(x_1, \ldots, x_{n_1} \mid z_1, \ldots, z_{n_3})F_{\mathbf{Y}|\mathbf{Z}}(y_1, \ldots, y_{n_2} \mid z_1, \ldots, z_{n_3}), \end{aligned}$$

where $F_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}$, $F_{\mathbf{X}|\mathbf{Z}}$, $F_{\mathbf{Y}|\mathbf{Z}}$ are the joint and marginal cumulative distribution functions (CDF) of $\mathbf{X}, \mathbf{Y}$ conditional on $(Z_1, \ldots, Z_{n_3}) = (z_1, \ldots, z_{n_3})$. This may be written $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}$.

We also allow $\mathbf{Z} = \{\}$, that is, $n_3 = 0$. In this case, $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}$ implies that $\mathbf{X}$ and $\mathbf{Y}$ are *unconditionally* independent, formally,

$$F_{\mathbf{X},\mathbf{Y}}(x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) = F_{\mathbf{X}}(x_1, \ldots, x_{n_1})F_{\mathbf{Y}}(y_1, \ldots, y_{n_2}),$$

where $F_{\mathbf{X},\mathbf{Y}}$, $F_{\mathbf{X}}$, $F_{\mathbf{Y}}$ are the unconditional joint and marginal cumulative distribution functions (CDF) of $\mathbf{X}, \mathbf{Y}$. This may be written $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \{\}$, or simply $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$.

We finally note that we may say, for example, that A and B are conditionally independent *given* C; or equivalently, that A and B are independent *conditional* on C (when $(A \perp\!\!\!\perp B) \mid C$). ///

We offer a few examples illustrating conditional independence.

Example 2.6 Suppose N is a positive random integer. Once N is observed, it is considered fixed, then X, Y are sampled independently from a binomial distribution with probability parameter p and sample size N.

When we say X, Y are independent, once N is considered fixed, we mean $(X \perp\!\!\!\perp Y) \mid N$. But if we do not condition on N, X and Y are *not* independent. Suppose we do not know the value of N. Then an observation of X gives us *some* information about N. At the very least, we would know that $N \geq X$. This in turn gives us information about Y. So X and Y are not independent unless we condition on N. □

Example 2.7 A dice is tossed independently three times. Let S_1, S_2, S_3 be the cumulative totals. For example, if the three outcomes are, in order, 3, 1, 5, then $S_1 = 3$, $S_2 = 4$ and $S_3 = 9$. Clearly, S_1 and S_3 are not independent. For example, the reader may verify the counter-example:

$$\begin{aligned} P(S_3 = 18 \mid S_1 = 6) &= 1/36, \text{ but} \\ P(S_3 = 18 \mid S_1 = 5) &= 0. \end{aligned}$$

On the other hand $(S_1 \perp\!\!\!\perp S_3) \mid S_2$. Once S_2 is known, the distribution of S_3 will not depend on the outcome of S_1. □

2.2.3 Conditional independence and the Bayesian network model

What defines a Bayesian Network? Despite the term *graphical model*, the *graph* need not be the most important object defining the Bayesian network model. It is sometimes helpful to think of this model primarily as a type of joint distribution $g = g(x_1, x_2, \ldots, x_n)$ of random variables $X_1, X_2, \ldots, X_n$ that satisfies certain constraints imposed by a graph G.

Then what role is played by the graph? A joint distribution g does not define a Bayesian network unless it satisfies certain constraints. Those constraints are imposed by a graph G (a DAG, to be precise) which has n nodes labeled by the random variables $X_1, X_2, \ldots, X_n$. Moreover, these constraints take the form of conditional independence statements. A DAG, therefore may be equivalently thought of as a list of conditional independence statements.

Example 2.8 The DAG in Figure 2.4 imposes the indicated list of conditional independence statements (this list is not exhaustive). The rules governing this will be discussed below. □

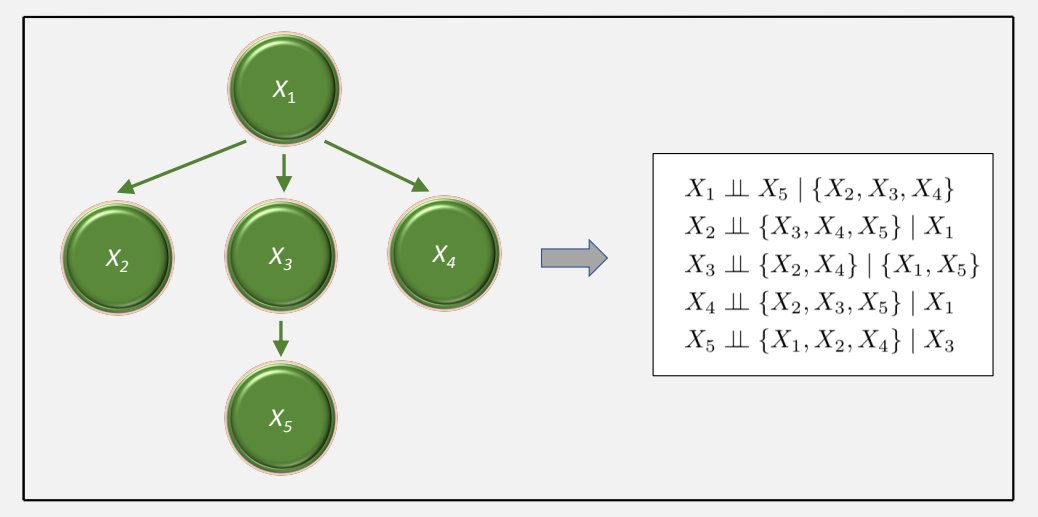

Figure 2.4: DAG used in Example 2.8. Conditional independence statements generated by the DAG are shown to the right.

So far, our discussion has been purely descriptive, so the following questions must be answered next:

QUESTION 1 According to what rules does a DAG such as the one shown in Figure 2.4 generate a set of conditional independence statements?

QUESTION 2 How can we construct a joint distribution $g(x_1, x_2, \ldots, x_n)$ which is constrained to conform to a specific set of conditional independence statements?

We next discuss in Sections 2.2.4 and 2.2.5 two methods which provide an answer to the first question. The second question will be answered in Section 2.3.

2.2.4 Markov blankets

The *Markov blanket* provides one method by which conditional independence statements are imposed by a DAG. The key is in the following definition.

Definition 2.4 Given a DAG G, the *Markov blanket* of a node j is the union of

(a) The set P_j of all parents of node j;
(b) The set C_j of all children of node j;
(c) The set of all parents of children of node j, other than node j itself.

Denote this set of nodes B_j. ///

Note that a more general concept of the Markov blanket exists for a broader class of probabilistic graphical models (see Example 2.14 below). Definition 2.4 gives the definition of a Markov blanket for Bayesian networks. The definition of a Markov blanket may differ for other types of graphical models. See, for example, Koller and Friedman (2009).

Conditional Independence Statements [Markov Blankets] The rules for generating the conditional independence statements for the DAG of Figure 2.4 can now be stated. Let V be the set of all nodes. Using the notation of Definition 2.4, the conditional independence statements are then

$$(\{X_j\} \perp\!\!\!\perp V - \{X_j\} - B_j) \mid B_j, \;\; j = 1, \ldots, n. \tag{2.2}$$

In other words, each node is conditionally independent of all nodes *outside* its Markov blanket, given the Markov blanket.

Example 2.9 Recall Example 1.5 (Chemical Reaction - Version 2). Suppose, to fix ideas, the correct model can be represented graphically as

$$a \to c \to b.$$

We can recognize this as a DAG. Furthermore, b has parent c, no children, and shares no children with another node. The Markov blanket of b is therefore $B_b = \{c\}$ (Definition 2.4), and the following conditional independence statement is imposed:

$$(a \perp\!\!\!\perp b) \mid c.$$

Clearly, a and b will be dependent, but transiently so, being independent conditionally on c. Given our understanding of chemical reactions, we might conclude that c, and not a, is the direct cause of b. □

Example 2.10 Consider node X_3 of Figure 2.4. The parent set is $P_3 = \{X_1\}$. The child set is $C_3 = \{X_5\}$. There are no other parent of children in C_3. So $B_3 = \{X_1, X_5\}$. Then $V = \{X_1, X_2, X_3, X_4, X_5\}$, so this leads to conditional independence statement

$$(\{X_3\} \perp\!\!\!\perp V - \{X_3\} - B_3) \mid B_3, \text{ equivalent to } (X_3 \perp\!\!\!\perp \{X_2, X_4\}) \mid \{X_1, X_5\}.$$

□

2.2.5 *D*-separation

Markov blankets provide an intuitive way of deriving conditional independence statements from a DAG, and they make clear the connection between Bayesian networks and Markov chains. However, we will see that they will

not exhaust all relevant conditional independence statements. A more comprehensive method involves the idea of *d-separation* (Pearl *et al.*, 1989; Koller and Friedman, 2009). The advantage of this method is that it can be used to test whether or not any putative conditional independence statement $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}$ is imposed by the DAG.

Definition 2.5 (D-separation) Suppose a and b are distinct nodes of a DAG. Let L be an *arc* between a and b. Then a node v on L is a *collider* if two edges on L are directed towards it (note that neither a nor b can be a collider on L). Let C be a subset of nodes. Then the arc L is *active* given C if both of the following rules hold:

Rule 1: If z is any collider on L, then either z or one of its descendants is in C.

Rule 2: If z is on L and is not a collider, then it is not in C.

Conversely, we say that an *arc* between a and b is *blocked* by C if at least one of the two rules does not hold (that is, the arc L is *not* active given C).

Then let A, B, C be three disjoint subsets of nodes. We say C *d-separates* A and B if there are no arcs joining any pair of nodes $a \in A$, $b \in B$ which are *active* given C. If C does not *d-separate* A and B, then C *d-connects* A and B. ///

Conditional Independence Statements [D-separation] The rule for generating conditional independence statements via d-separation is now simply:

$$\text{If } C \text{ } d\text{-separates } A \text{ and } B, \text{ then } (A \perp\!\!\!\perp B) \mid C. \tag{2.3}$$

Example 2.11 We will repeat Example 2.9 using d-separation. To do this, we need to show that $\{c\}$ d-separates $\{a\}$ and $\{b\}$. There is only one arc L from a to b, namely $a \to c \to b$. This arc contains a node which is not a collider, but is in $\{c\}$. Therefore, by Definition 2.4, L is *not* active given $\{c\}$, $\{c\}$ d-separates $\{a\}$ and $\{b\}$, and the conditional independence statement $(a \perp\!\!\!\perp b) \mid c$ is imposed by the DAG. □

Example 2.12 Consider the DAG of Figure 2.1.

(a) Let $A = \{X_6\}$ and $B = \{X_1, X_2, X_3\}$. What set of nodes C d-separates A and B? Let $C = \{X_4\}$. Every arc from a node in A to a node in B passes through X_4. Is X_4 a collider? Although this node has in-degree 2 (two edges are directed to X_4), a node is defined as a collider only with respect to a specific arc. Then it is easily verified that X_4 is *not* a collider on any arc joining any node in A to any node in B, so Rule

2 of Definition 2.5 is violated, since $X_4 \in C$. Thus, since there are no arcs joining $a \in A$ and $b \in B$ which are active given C, we conclude that C d-separates A and B, which imposes conditional independence statement

$$(\{X_6\} \perp\!\!\!\perp \{X_1, X_2, X_3\}) \mid \{X_4\}.$$

(b) Let $A = \{X_5\}$ and $B = \{X_1, X_2, X_3, X_4\}$. What set of nodes C d-separates A and B? Let $C = \{\}$, the empty set (this is a valid set for this type of analysis). Every arc L from a node in A to a node in B passes through X_6, which will be a collider. But neither X_6, nor any of its descendants, is in C. Thus, since there are no arcs joining $a \in A$ and $b \in B$ which are active given C, we conclude that C d-separates A and B, which imposes conditional independence statement

$$(\{X_5\} \perp\!\!\!\perp \{X_1, X_2, X_3, X_4\}) \mid \{\},$$

or, equivalently

$$\{X_5\} \perp\!\!\!\perp \{X_1, X_2, X_3, X_4\}.$$

□

Remark 2.1 There are a few comments regarding Example 2.12 worth making.

- Regarding Example (a), we note that node X_4 is a collider on *some* path (for example, $X_1 \rightarrow X_2 \rightarrow X_4 \leftarrow X_3$), but is not a collider on any of the paths joining nodes in A to nodes in B.
- Regarding Example (b), it is worth drawing attention to the peculiar wording required to make the argument. The statement

  `But neither` X_6`, nor any of its descendants, is in` C `...`

  is true because X_6 has no descendants, and C does not contain any nodes.

///

The conditional independence structure imposed by a DAG might sometimes seem complex and unintuitive. A careful study of the concepts of Markov blankets and d-separation is therefore needed for a thorough understanding of the Bayesian network model.

Example 2.13 Consider the DAG of Figure 2.5. If we set $A = \{X_1\}$, $B = \{X_3\}$, then for what sets of nodes C does the DAG impose the conditional independence statement $(A \perp\!\!\!\perp B) \mid C$?

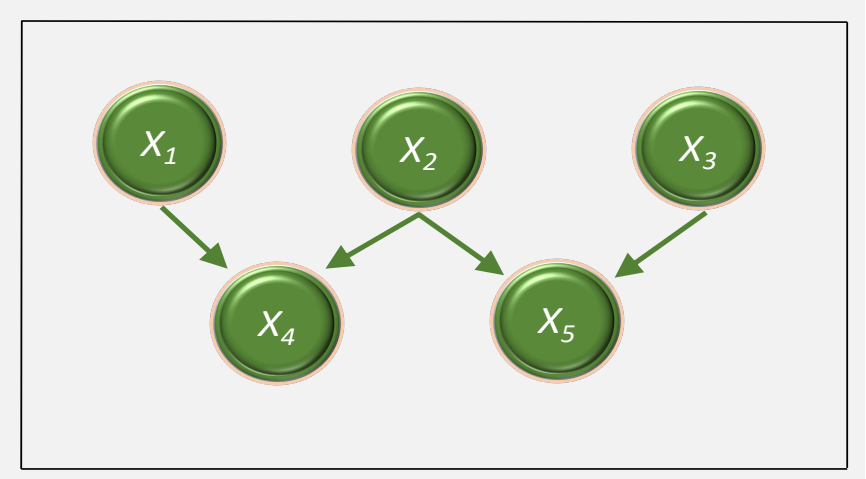

Figure 2.5: DAG used in Example 2.13.

(a) First note the child set $C_1 = \{X_4\}$, and the one child of node X_1, namely X_4, also has parent X_2. By Definition 2.4 the Markov blanket of X_1 is therefore $B_1 = \{X_2, X_4\}$. This imposes conditional independence statement

$$(X_1 \perp\!\!\!\perp X_3) \mid \{X_2, X_4\}.$$

(b) A similar argument shows that the Markov blanket of X_3 is $B_3 = \{X_2, X_5\}$, so the DAG also imposes conditional independence statement

$$(X_1 \perp\!\!\!\perp X_3) \mid \{X_2, X_5\}.$$

(c) Then set $C = \{X_4\}$. Does C d-separate A and B? There is only one arc joining X_1 and X_3. There is a collider, X_5, which is not in C, and which has no descendants. Therefore, Rule 1 of Definition 2.5 is violated, the arc is *not* active given C, and we conclude that C d-separates A and B. The DAG therefore imposes conditional independence statement

$$(X_1 \perp\!\!\!\perp X_3) \mid \{X_4\}.$$

(d) Next, set $C = \{X_5\}$. Does C d-separate A and B? Essentially the same argument used to show that $\{X_4\}$ d-separates A and B can be used to show that $\{X_5\}$ also d-separates A and B. The DAG therefore imposes conditional independence statement

$$(X_1 \perp\!\!\!\perp X_3) \mid \{X_5\}.$$

(e) Next, set $C = \{X_4, X_5\}$. Does C d-separate A and B? There are two colliders on the arc joining X_1 and X_3, and both are in C, so that Rule 1 of Definition 2.5 is satisfied. There is an additional node on the arc

which is not a collider and which is not in C, so that Rule 2 is satisfied. Thus, the arc is active given C, and we conclude that C does *not* d-separate A and B. For more on this example, see Example 2.16 below.

(f) Finally, set $C = \{X_2, X_4, X_5\}$. Does C d-separate A and B? There is only one arc joining X_1 and X_3. Rule 2 of Definition 2.5 is violated, since the arc joining X_1 and X_3 contains a node, X_2, which is not a collider, but which is in C (although Rule 1 still holds). Thus, the arc is *not* active given C, and we conclude that C d-separates A and B. The DAG therefore imposes conditional independence statement

$$(X_1 \perp\!\!\!\perp X_3) \mid \{X_2, X_4, X_5\}.$$

□

Example 2.14 (Relationship of Markov Blankets to D-separation) The Markov blanket can be defined in terms of d-separation. In particular, following Definition 2.4, it may be shown that B_j is the minimal set of nodes which d-separates node j from all remaining nodes of a DAG. This means that the set of all conditional independence statements generated by Markov blankets (Equation (2.2)) is a strict subset of those generated by d-separation (Equation (2.3)).

See Pearl (1988), as well as Section 2.1.4 of Nagarajan *et al.* (2013) for a more informal discussion of Markov blankets. □

2.3 Formal Definition of the Bayesian Network Model

So far, we have discussed how a DAG can imply a collection of conditional independence statements. The next problem is to develop a method of constructing joint densities for the nodes which conform to those statements. The theory of Bayesian networks provides a rigorous solution to this problem. However, it worth considering first how this may be done in an informal but intuitive way. We do so in the next example.

Example 2.15 (Constructing a Bayesian network) In this example we will attempt to build directly a model which conforms to the set of conditional independence statements imposed by the DAG in Figure 2.4. To do this we will follow the graph's sequential structure. Let $\epsilon_1, \ldots, \epsilon_5$ be five independent random variables, of any kind. Consider the following rules:

Rule 1: Node X_1 is the "first" node in what appears to be a sequential process. Set

$$X_1 = \epsilon_1.$$

Rule 2: Each of the remaining nodes have exactly one parent, say p_j. The rule is simple. Each node inherits the value of its parent, plus an independent noise term. That is,

$$X_j = X_{p_j} + \epsilon_j,$$

where p_j is the parent of node X_j.

If we apply Rules 1 and 2, we have the following system of linear equations:

$$\begin{aligned} X_1 &= \epsilon_1 && = \epsilon_1 \\ X_2 &= X_1 + \epsilon_2 && = \epsilon_1 + \epsilon_2 \\ X_3 &= X_1 + \epsilon_3 && = \epsilon_1 + \epsilon_3 \\ X_4 &= X_1 + \epsilon_4 && = \epsilon_1 + \epsilon_4 \\ X_5 &= X_3 + \epsilon_5 && = \epsilon_1 + \epsilon_3 + \epsilon_5. \end{aligned}$$

Does this model satisfy the conditional independence statements listed in Figure 2.4? To answer this question, remember that when we condition on a random variable, we are regarding its value as fixed, or constant. We will adopt a notational device to represent this, putting in square brackets any random variable or expression on which we are conditioning, and which we therefore wish to regard as fixed. For example, if we condition on X_1, we replace X_1 in any expression with $[X_1]$, or alternatively, ϵ_1 with $[\epsilon_1]$.

(a) Consider the conditional independence statement:

$$(X_2 \perp\!\!\!\perp \{X_3, X_4, X_5\}) \mid X_1. \tag{2.4}$$

We are conditioning on X_1, so write, following the preceding equations:

$$\begin{aligned} X_2 &= [X_1] + \epsilon_2 \\ X_3 &= [X_1] + \epsilon_3 \\ X_4 &= [X_1] + \epsilon_4 \\ X_5 &= X_3 + \epsilon_5 = [X_1] + \epsilon_3 + \epsilon_5. \end{aligned}$$

Once we condition on X_1, $[X_1]$ is interpreted as a constant. Then X_2 depends only on ϵ_2. The remaining nodes X_3, X_4, X_5 depend exclusively on $\epsilon_3, \epsilon_4, \epsilon_5$, which are independent of ϵ_2. So the conditional independence statement of Equation (2.4) holds.

(b) Consider the conditional independence statement:

$$(X_5 \perp\!\!\!\perp \{X_1, X_2, X_4\}) \mid X_3. \tag{2.5}$$

As in the previous example, write:

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \epsilon_1 + \epsilon_2 \\ X_4 &= \epsilon_1 + \epsilon_4 \\ X_5 &= [X_3] + \epsilon_5. \end{aligned}$$

Once we condition on X_3, $[X_3]$ is interpreted as a constant. Then X_5 depends only on ϵ_5. The remaining nodes X_1, X_2, X_4 depend exclusively on $\epsilon_1, \epsilon_2, \epsilon_4$, which are independent of ϵ_5. So the conditional independence statement of Equation (2.5) holds.

□

Example 2.16 (Example 2.13 Continued) We can use the approach of Example 2.15 to understand why, in Example 2.13, the conditional independence statement $(X_1 \perp\!\!\!\perp X_3) \mid \{X_4\}$ holds but $(X_1 \perp\!\!\!\perp X_3) \mid \{X_4, X_5\}$ does not (parts (c) and (e)). Using the two rules of Example 2.15 would give here

$$\begin{aligned} X_4 &= X_1 + X_2 + \epsilon_4, \\ X_5 &= X_2 + X_3 + \epsilon_5, \end{aligned} \tag{2.6}$$

where ϵ_4, ϵ_5 are independent random variables associated with nodes 4 and 5. We can express the joint distribution conditional on $\{X_4, X_5\}$ by setting $[X_4] = s$ and $[X_5] = t$ for two fixed constants s, t. This imposes the two linear constraints

$$\begin{aligned} [X_4] &= [X_1 + X_2 + \epsilon_4] = s, \\ [X_5] &= [X_2 + X_3 + \epsilon_5] = t. \end{aligned}$$

At this point, it is instructive to subtract $[X_5]$ from $[X_4]$, noting that the term X_2 will cancel, which results in the following constraint:

$$(X_1 + \epsilon_4) - (X_3 + \epsilon_5) = s - t. \tag{2.7}$$

How can we interpret Equation (2.7)? We can take $s - t$ to be constant, and then interpret (2.7) as a "noisy" linear constraint on the random variables X_1 and X_3, with ϵ_4, ϵ_5 playing the role of "noise". We lose no generality by making the variance of ϵ_4, ϵ_5 as small as we like, and so we can accept, approximately, the linear constraint:

$$X_1 - X_3 \approx s - t. \tag{2.8}$$

It is easily verified that two independent random variables are no longer independent when conditioned on a constraint such as Equation (2.8).

On the other hand, the conditional independence statement $(X_1 \perp\!\!\!\perp X_3) \mid \{X_4\}$ holds, since X_3 does not appear in the constructive definition of X_1 or X_4 (Equation (2.6)). □

Example 2.15 suggests that the construction of a joint distribution which conforms to the conditional independence statements imposed by a DAG largely involves mimicking its flow in some say. In the next section, we will see that this is the case.

2.3.1 Factorization and the local and global Markov properties

Recall that *any* joint density of random variables $X_1, \ldots, X_n$ can be decomposed in the following way:

$$\begin{aligned} g(x_1, x_{n+2}, \ldots, x_n) &= g(x_n \mid x_{n-1}, \ldots, x_1) \times g(x_{n-1}, \ldots, x_1) \\ &= g(x_n \mid x_{n-1}, \ldots, x_1) \times g(x_{n-1} \mid x_{n-2}, \ldots, x_1) \times g(x_{n-2}, \ldots, x_1) \\ &= \left(\prod_{j=1}^{n-1} g(x_{n-j+1} \mid x_{n-j}, \ldots, x_1) \right) \times g(x_1). \end{aligned}$$

This is referred to as the *chain rule*. This expression simplifies considerably for a Markov chain, since by the memoryless property we have

$$g(x_j \mid x_{j-1}, \ldots, x_1) = g(x_j \mid x_{j-1}),$$

and so we have the much simpler form

$$g(x_1, x_{n+2}, \ldots, x_n) = \left(\prod_{j=1}^{n-1} g(x_{n-j+1} \mid x_{n-j}) \right) \times g(x_1). \tag{2.9}$$

This type of simplification also occurs for a Bayesian network, which will be built from conditional densities of the form $g(x_j \mid P_j)$, interpretable as the distribution of node X_j conditional on that node's parents.

Example 2.17 For the DAG in Figure 2.1 the conditional distribution

$$g(x_6 \mid P_6) = g(x_6 \mid x_4, x_5)$$

will play an important role. It would give, for example, the probability of missing work for an individual who has acquired an infection, and has resistance to antibiotics. □

Example 2.18 (Founders) Founders are an important special case. Suppose X_1 is a founder. Then $P_1 = \{\}$, and we have

$$g(x_1 \mid P_1) = g(x_1 \mid \{\}) = g(x_1).$$

Here we are interpreting the distribution of a random variable conditional on an *empty* collection of random variables as the *unconditional* distribution. It turns out that this convention may be applied generally. □

A Markov chain (of a finite number of transitions) is also a Bayesian network, representable by the DAG shown in Figure 2.6. Then note that

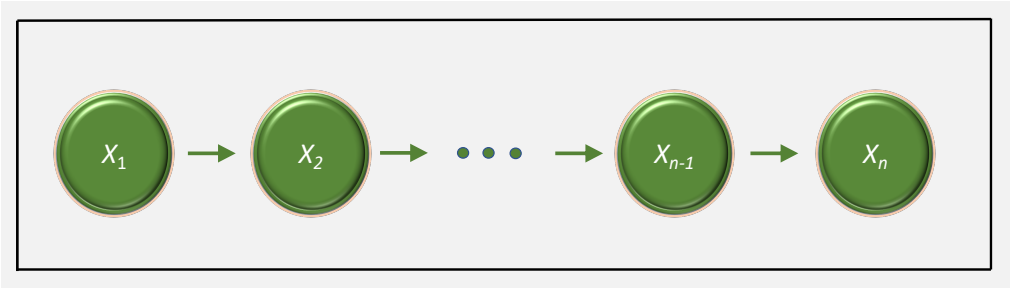

Figure 2.6: DAG representation of a Markov Chain.

Equation (2.9), which gives the joint distribution of a Markov chain, can be rewritten using the parent sets:

$$g(x_1, x_{n+2}, \ldots, x_n) = \prod_{j=1}^{n} g(x_j \mid P_j), \tag{2.10}$$

noting that the term $g(x_1)$ appearing in Equation (2.9) is represented in Equation (2.10) as

$$g(x_1 \mid P_1) = g(x_1 \mid \{\}) = g(x_1),$$

since X_1 is a founder in the representational DAG of Figure 2.6.

At this point, we have enough to formally define a Bayesian network supported by a rigorous mathematical foundation. We first define the method by which Bayesian networks may be constructed.

Definition 2.6 (Factorization over a DAG) Let $\mathbf{X} = (X_1, \ldots, X_n)$ be n random variables which label n nodes of a DAG, say $G = (V, E)$. We say a joint distribution g on $\mathbf{X}$ can be *factorized* over G if it may be expressed in the following way:

$$g(x_1, x_2, \ldots, x_n) = \prod_{j=1}^{n} g(x_j \mid P_j), \tag{2.11}$$

where P_j is the parent set of node X_j according to the DAG G. ///

The essential point of the theory of Bayesian networks is that a distribution which factorizes over a DAG G also satisfies the conditional independence statements imposed by G. This idea can be expressed using the *local* and *global Markov properties*.

Definition 2.7 (Local and Global Markov Properties) Let $\mathbf{X} = (X_1, \ldots, X_n)$ be n random variables which label n nodes of a DAG, say $G = (V, E)$.

We say a joint distribution g on $\mathbf{X}$ satisfies the *local Markov property* with respect to G if it satisfies all conditional independence statements of the form:

$$(X_j \perp\!\!\!\perp D_j^c - \{X_j\}) \mid P_j, \;\; j = 1, \ldots, n, \tag{2.12}$$

where P_j is the set of all parents of node j, and $D_j^c - \{X_j\}$ is the set of all non-descendants of node j, excluding node j itself.

We say a joint distribution g on $\mathbf{X}$ satisfies the *global Markov property* with respect to G if it satisfies all conditional independence statements of the form:

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}, \tag{2.13}$$

where $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are disjoint sets of node labels such that $\mathbf{X}$ and $\mathbf{Y}$ are nonempty, and $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$. ///

Remark 2.2 Note that in the conditional independence statement of Equation (2.12), the set D_j^c includes the parents P_j of node j. This is not a contradiction. When we condition on the random variables in P_j, they become constant in the resulting distribution. Formally, a constant is independent of any other random variable (similarly, for two events A, B, if $P(A) = 1$, then $P(AB) = P(B) = 1 \times P(B) = P(A)P(B)$).

However, for this reason some texts express the conditional independence statement (2.12) in the equivalent form

$$(X_j \perp\!\!\!\perp D_j^c - \{X_j\} - P_j) \mid P_j, \;\; j = 1, \ldots, n,$$

for greater clarity. ///

Example 2.19 (The Global Property Implies the Local Property) A good question to ask at this point might be why both the local and global Markov properties are needed. We will discuss this later in this section, but we first show that the global property implies the local property, or equivalently, that a conditional independence statement of the form (2.12) is simply a special case of the class of conditional independence statements of the form (2.13) (compare this to Example 2.14).

In turn, this may be shown by verifying that a node j is d-separated from any non-parental, non-descendant k by the parent set P_j. Clearly, any arc L joining nodes j and k must pass though a node from the parent set P_j. Therefore, if $P_j = \{\}$, then there are no arcs joining j and k, in particular, there are no *active* arcs joining j and k, so by Definition 2.5 j and k are d-separated by P_j.

Next, assume that P_j is not empty. Suppose node $z \in P_j$ is on L. The edge joining z and j points to j, so z cannot be a collider on L. Rule 2 of Definition 2.5 is violated, so L is not active given P_j. This is true of all arcs joining nodes j and k, so we may conclude that j and k are d-separated by P_j. □

The global Markov property (Definition 2.7) and the factorization property (Definition 2.6) are shown to be equivalent (with an important caveat) in the following theorem:

Theorem 2.1 (Factorization Theorem) Let $\mathbf{X} = (X_1, \ldots, X_n)$ be n random variables which label n nodes of a DAG, say $G = (V, E)$. Let g be the joint distribution of $\mathbf{X}$. Then

(i) The following three statements are equivalent:
   (A) The distribution g can be factorized over G (Definition 2.6);
   (B) The distribution g satisfies the local Markov property with respect to G (Definition 2.7);
   (C) The distribution g satisfies the global Markov property with respect to G (Definition 2.7).

(ii) Suppose $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are disjoint sets of node labels such that $\mathbf{X}$ and $\mathbf{Y}$ are *not* d-separated by $\mathbf{Z}$, with respect to G. Then there exists a distribution which factorizes over G, but for which the conditional independence statement $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}$ does *not* hold.

///

Remark 2.3 Theorem 2.1 follows from, for example, Theorems 3.1-3.4 of Koller and Friedman (2009) or Theorem 5.14 of Cowell *et al.* (2006). ///

Remark 2.4 We consider the importance of Part (ii) of Theorem 2.1. Example 2.19 makes clear that the conditional independence statements of Equation (2.12) enumerated by the local Markov property are a strict subset of those enumerated by the global Markov property (Equation (2.13)).

However, both the local and global Markov properties suffice to characterize a joint distribution g which can be factorized over G, as stated by Part (i) of Theorem 2.1. Each formulation has its advantages. The local Markov

property is the simpler, more intuitive and more compact representation, which has obvious advantages in mathematical analysis.

On the other hand, the global Markov property comes closer to defining an exhaustive list of conditional independence statements needed to characterize the Bayesian network. However, the strongest possible statement of this equivalence does not hold. In particular, it is *not* true that if g factorizes over G, it follows that $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}$ holds if and only if $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$. A counter-example is easy to construct. If all node labels $X_1, \ldots, X_n$ are mutually independent, then *all possible* conditional independence statements are satisfied by g, yet g will factorize over any DAG G.

The wording of Part (ii) of Theorem 2.1 needs to be carefully considered. Assume g factorizes over G. If $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$, then $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}$ must hold. If $\mathbf{Z}$ does not d-separates $\mathbf{X}$ and $\mathbf{Y}$, then $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}$ may hold, but there will be some g' which factorizes over G for which $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}$ does not hold. In other words, $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ if and only if $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \mid \mathbf{Z}$ holds for *all* g which factorize over G. ///

This formally defines the Bayesian network model:

Definition 2.8 (Bayesian Network) Given a set of nodes $V = \{1, \ldots, n\}$ labeled by random variables $X_1, \ldots, X_n$, a Bayesian network is a pair (G, g), where G is a DAG defined on nodes V, P_j is the parent set of node j with respect to G, and $g = g(x_1, x_2, \ldots, x_n)$ is a joint density for the random variable labels which factors according to Equation (2.11).

In this case, g satisfies the local and global Markov properties of Definition 2.7. ///

The mathematical justification for Definition 2.8 is described in, for example, Pearl *et al.* (1989) and is summarized by Theorem 2.1 above.

2.3.2 Parametric models and estimation

The inference of Bayesian networks can be decomposed into subproblems of one of two types:

Problem 1: Estimation of DAG G;

Problem 2: Estimation of the conditional distributions $g(x_i \mid P_j)$, $j = 1, \ldots, n$, which define the factorization of Equation (2.11).

From the point of view of statistical theory, Problem 1 concerns the inference of conditional independence statements (those imposed by G). This is, in fact, a well defined inference problem. Tests for conditional independence have been proposed as early as 1924 by R. A. Fisher (Fisher, 1924).

Problem 2 concerns the form of the conditional distributions $g(x_j \mid P_j)$, which describe the probabilistic relationships implied by the edges. So far, we have been primarily concerned with the DAG G. This is entirely appropriate, since it is the DAG that gives the Bayesian network model its distinguishing properties. However, the conditional distributions $g(x_j \mid P_j)$ will usually depend on unknown parameters which must be estimated using the same data used to estimate G. This is a significant estimation problem, which will be discussed in future chapters. For now, the next example will give a sense of what this involves.

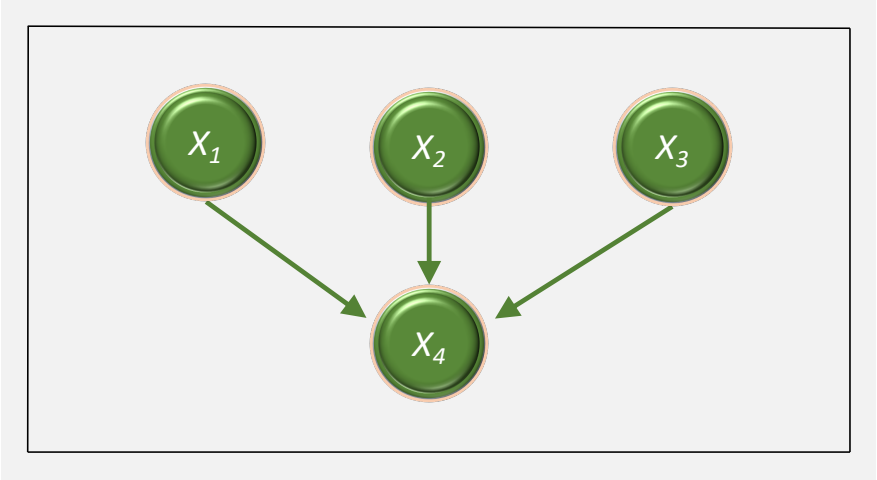

Figure 2.7: DAG used in Example 2.20 (Gaussian Bayesian network).

Example 2.20 (Gaussian Bayesian Networks) In a Bayesian network model, the form of the conditional distributions $g(x_j \mid P_j)$ of Equation (2.11) will depend on the properties of the random variables $X_1, \ldots, X_n$ used to label the nodes of the DAG G. These may be continuous, discrete, or some combination of the two. When all X_i are approximately normally distributed (perhaps after a logarithmic transformation), a *Gaussian Bayesian network* is commonly used, and $(X_1, \ldots, X_n)$ becomes a multivariate normal random vector constrained by the conditional independence statements imposed by G.

One significant advantage of the Bayesian network model is that it simplifies the construction of joint densities of potentially very high dimension (see also Example 2.21 below). Consider the DAG in Figure 2.7. Suppose X_1, X_2, X_3, X_4 are normally distributed. We saw in Example 2.15 how the conditional independence statements imposed by a DAG can be captured by the simple device of building a functional relationship between a node X_j and its parents P_j. In that example each node had no more than one parent,

but the method is much the same when this is not the case. For example, for the DAG in Figure 2.7 this relationship could be expressed

$$X_4 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_4, \tag{2.14}$$

where ϵ_4 is an independent normal random variable with mean zero; and $\beta_0, \ldots, \beta_3$ are constant but unknown coefficients to be estimated using the available data. Equation (2.14) can be recognized as a multiple linear regression model with dependent variable X_4 and independent variables X_1, X_2, X_3. This means the coefficients β_j can be estimated using any conventional statistical software. Furthermore, this procedure remains the same when the DAG is a subgraph of a larger graph in which X_4 retains the same parent set $P_4 = \{X_1, X_2, X_3\}$. □

We next introduce another type of Bayesian network model which will be discussed in later chapters.

Example 2.21 (Dependence Trees) It is worth noting a type of precursor of the Bayesian network described in Chow and Liu (1968), referred to as a *dependence tree*. It has the same type of conditional independence structure as a Bayesian network, but this is expressed using a *spanning tree*, which is an undirected graph in which all nodes are connected by a path, but which contains no cycles. It may be shown that a dependence tree is equivalent to a Bayesian network associated with a DAG in which no node has more than one parent. We will see in later chapters that this constraint results in considerable computational efficiency. □

2.3.3 Model identifiability

So far, we have seen:

(1) Various methods of determining conditional independence statements imposed by a DAG (Markov blankets and d-separation);

(2) A method of constructing a joint distribution on the nodes of a DAG which conforms to those conditional independence statements (the Factorization Theorem).

Clearly, there is a very important third step. The motivation for using graphical models is typically the insight into a process offered by the graph. But we have claimed that the Bayesian network model may not be able to identify a single graph as correct (for example, see Example 1.5). It is important to emphasis that we are not referring to the statistical error inevitable in any inference. We are referring to something more fundamental.

Definition 2.9 (Identifiability and Consistency) Suppose we are given a class of putative models indexed by Θ. For each $\theta \in \Theta$ there exists a distribution g_θ for a random vector $\mathbf{X}$. Let $d(\theta_1, \theta_2) \geq 0$ be a distance function on Θ, such that $d(\theta_1, \theta_2) = 0$ if and only if $\theta_1 = \theta_2$. Suppose θ^* is the true model, and let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be an unbounded sample from distribution g_{θ^*}. We say the model is *identifiable* if there exists a sequence of estimators $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$, $n \geq 1$, such that $\lim_{n\to\infty} P(d(\hat{\theta}_n, \theta^*) < \epsilon) = 0$ for all $\epsilon > 0$ and $\theta^* \in \Theta$. We then say such an estimator is *consistent*. ///

The reader may wish to consult a good introduction to the theory of statistical inference for more on the concept of *consistency*, for example Hogg *et al.* (2018) or Casella and Berger (2002).

In our case Θ would be the class of Bayesian network models on a given set of nodes. A parameter $\theta \in \Theta$ might specify the DAG G, but also additional parameters (which we will call *auxiliary parameters*) defining the conditional distributions used in the factorized distribution of Equation (2.11) (see Example 2.20).

To understand the issue of identifiability as it relates to Bayesian networks, two facts are crucial:

(1) Conditional independence statements can be consistently tested, and conditional distributions $g(x_j \mid P_j)$ can be consistently estimated;

(2) However, the totality of conditional independence statements imposed by a DAG do not uniquely determine that DAG.

Thus, the lack of identifiability of the DAG G in a Bayesian network model is not related to the issue of consistency, and so can not be overcome with a large enough sample size. The conditional distributions $g(x_i \mid P_j)$ and the conditional independence statements can be consistently estimated. Thus, the entire joint distribution g defining a Bayesian network (G, g) can be consistently estimated. The issue is that multiple DAGs may impose exactly the same conditional independence statements. These, in turn, will lead to equivalent factorizations (Equation (2.11)), although this may not be readily apparent.

To make this idea precise, we need the concept of *equivalence classes*, which we discuss next.

2.4 Equivalence Classes

Recall that in Example 1.5 (Chemical Reaction - Version 2) there was interest in the order in which three agents a, b, c acted in a chemical reaction. Of

course, there are six possible hypotheses:

$$a \to b \to c$$
$$a \to c \to b$$
$$b \to a \to c$$
$$b \to c \to a$$
$$c \to a \to b$$
$$c \to b \to a.$$

It was claimed in that example that a Bayesian network model (G, g) would be able to reduce the number of hypotheses from six to two (by identifying the middle agent), but would not be able to resolve those final two. For example, if c was identified as the middle agent, we would still be left with the problem of resolving the remaining hypotheses:

$$a \to c \to b$$
$$b \to c \to a.$$

Clearly, we need to understand to what degree the Bayesian network model is able to resolve multiple hypotheses concerning G. The notion of *equivalence classes* is central to this problem.

QUESTION Suppose we can consistently estimate all conditional independence statements and auxiliary parameters. Does this imply that the Bayesian network model is identifiable?

ANSWER The density of the Bayesian network given in Equation (2.11) of the Factorization Theorem can be consistently estimated. But because multiple DAGs can impose the same set of conditional independence statements, the underlying DAG itself is not in general identifiable.

It is worth examining this question in detail for a few specific models.

Example 2.22 To explore this issue, it helps to start with the simplest Bayesian network model that still retains some interesting structure. If we apply the Factorization Theorem, the joint density for (X_1, X_2) imposed by `DAG 1` of Figure 2.8 would be, using Equation (2.11),

$$\begin{aligned} g(x_1, x_2) &= g(x_1 \mid P_1) g(x_2 \mid P_2) \\ &= g(x_1 \mid \{\}) g(x_2 \mid x_1) \\ &= g(x_1) g(x_2 \mid x_1) \\ &= g(x_1, x_2). \end{aligned}$$

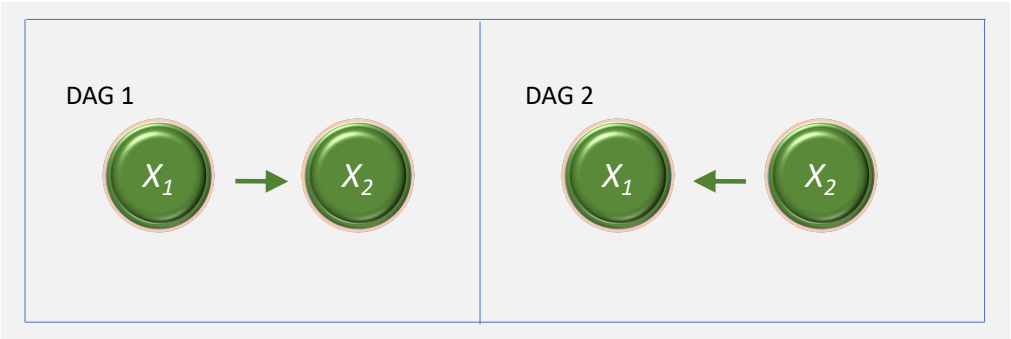

Figure 2.8: Examples of DAGs (two nodes and one edge).

What do we conclude from this? That any joint distribution on (X_1, X_2) is compatible with `DAG 1`. If we repeat the exercise for `DAG 2` of Figure 2.8, we similarly have

$$\begin{aligned} g(x_1, x_2) &= g(x_1 \mid P_1)g(x_2 \mid P_2) \\ &= g(x_1 \mid x_2)g(x_2 \mid \{\}) \\ &= g(x_1 \mid x_2)g(x_2) \\ &= g(x_1, x_2). \end{aligned}$$

The structure is the same as for `DAG 1`. Either model admits any form of dependence between X_1 and X_2, and are therefore not distinguishable. □

Clearly, we need to examine a more complex model to discern any variety of causal structure, and we only need to add one more node to do so.

Example 2.23 If we apply the Factorization Theorem to `DAG 1` of Figure 2.9 we have a joint distribution for (X_1, X_2, X_3) of the form

$$\begin{aligned} g(x_1, x_2, x_3) &= g(x_1 \mid P_1)g(x_2 \mid P_2)g(x_3 \mid P_3) \\ &= g(x_1 \mid x_2)g(x_2 \mid \{\})g(x_3 \mid x_2) \\ &= g(x_1 \mid x_2)g(x_2)g(x_3 \mid x_2). \end{aligned} \tag{2.15}$$

If we divide this expression by $g(x_2)$ we have the equivalent expression:

$$\frac{g(x_1, x_2, x_3)}{g(x_2)} = g(x_1, x_3 \mid x_2) = g(x_1 \mid x_2)g(x_3 \mid x_2). \tag{2.16}$$

It is not hard to verify that when this expression is compared to Definition 2.3 (of conditional independence) we may claim

$$(X_1 \perp\!\!\!\perp X_3) \mid X_2. \tag{2.17}$$

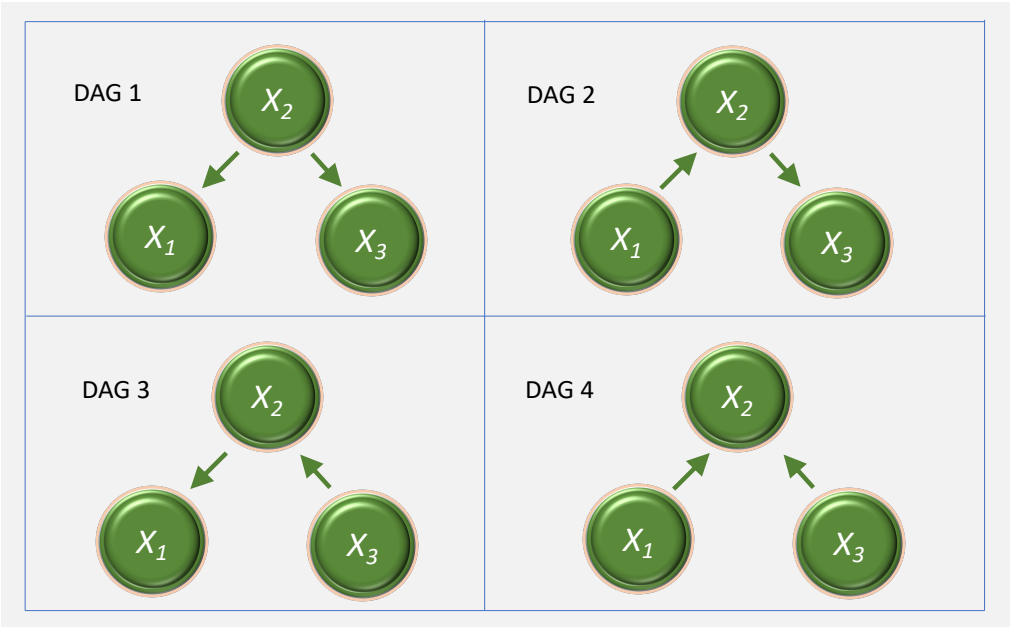

Figure 2.9: Examples of DAGs (three nodes and two edges).

We can reach the same conclusion by determining the Markov blanket of node X_1. For `DAG 1` we have $P_1 = \{X_2\}$, $C_1 = \{\}$, and node X_1 shares no children with other parents (Definition 2.4). The Markov blanket for node X_1 is therefore $B_1 = \{X_2\}$, so that `DAG 1` imposes the conditional independence statement

$$(X_1 \perp\!\!\!\perp X_3) \mid X_2,$$

which conforms to the same conclusion implied by the Factorization Theorem via Equation (2.16).

The structure of `DAG 2` and `DAG 3` is essentially the same as for `DAG 1`. The reader can verify that the Markov blanket for X_1 is also $B_1 = \{X_2\}$ for both, and a complete analysis would reveal that the conditional independence structure is exactly the same for `DAG 1`, `DAG 2` and `DAG 3`. In other words, we could not distinguish between them using observational data.

This leaves `DAG 4`. If we construct the joint distribution for (X_1, X_2, X_3) using the Factorization Theorem we obtain the form:

$$\begin{aligned} g(x_1, x_2, x_3) &= g(x_1 \mid P_1)g(x_2 \mid P_2)g(x_3 \mid P_3) \\ &= g(x_1 \mid \{\})g(x_2 \mid x_1, x_3)g(x_3 \mid \{\}) \\ &= g(x_1)g(x_2 \mid x_1, x_3)g(x_3). \end{aligned} \tag{2.18}$$

How is this related to the conditional independence statements imposed by `DAG 4`? We will first determine the Markov blanket for X_1, as we did for `DAG`

1, DAG 2 and DAG 3. Here, we have $P_1 = \{\}$ and $C_1 = \{X_2\}$. However, recall from Definition 2.4 that a Markov blanket of a node j includes

The set of all parents of children of node j, excluding node j.

X_2 is a child of X_1, X_3 is a parent of X_2, therefore X_3 is included in the Markov blanket of X_1, which is therefore $B_1 = \{X_2, X_3\}$. This differs from the Markov blanket of X_1 in DAG 1, DAG 2 and DAG 3 ($B_1 = \{X_2\}$), so we *cannot* use Markov blankets to test whether or not DAG 4 imposes the conditional independence statement $(X_1 \perp\!\!\!\perp X_3) \mid X_2$ given in (2.17).

Fortunately, we can use d-separation to resolve the matter. In particular, we need to determine whether or not the node subset $C = \{X_2\}$ d-separates X_1 and X_3. There is only one arc joining X_1 and X_3, namely $X_1 \rightarrow X_2 \leftarrow X_3$. Furthermore, on this arc X_2 is a collider, and is also contained in C. In addition, there are no other nodes on the arc which are *not* colliders (other than X_1 and X_3), so both Rule 1 and Rule 2 of Definition 2.5 hold, and the arc is active given C. This means X_2 d-connects X_1 and X_3, meaning that X_2 *does not* d-separate X_1 and X_3. So, we may conclude that DAG 4 *does not* impose the conditional independence statement $(X_1 \perp\!\!\!\perp X_3) \mid X_2$given in (2.17).

It is worth asking how the factorization of Equation (2.18) for DAG 4 compares to that of Equation (2.15), which will hold for DAG 1, DAG 2 and DAG 3. Note than any density for three variables may be decomposed as

$$g(x_1, x_2, x_3) = g(x_2 \mid x_1, x_3)g(x_1, x_3).$$

Therefore, the characteristic constraint imposed by the factorization of Equation (2.18) is precisely that

$$g(x_1, x_3) = g(x_1)g(x_3), \tag{2.19}$$

or, equivalently, that X_1 and X_3 are independent. Is this constraint imposed by DAG 4? Recall from the definition of conditional independence (Definition 2.3) that independence conditional on an empty set is equivalent to unconditional independence, which is implied for X_1, X_3 by Equation (2.19). Then DAG 4 imposes this constraint if X_1 and X_3 are d-separated by the empty set $C = \{\}$. As already noted, there is one arc joining X_1 and X_3, on which X_2 is a collider. Then consider Rule 1 of Definition 2.5. Neither X_2 nor any of its descendants is in C, so Rule 1 is violated, the arc is not active given C, and so X_1 and X_3 are d-separated by $C = \{\}$. This means $X_1 \perp\!\!\!\perp X_3$, as implied by Equation (2.19). See Example 2.12, Part (b), for further comment on this type of argument. □

2.4.1 Equivalence classes and v-structures

In the context of Bayesian networks, "causality" is a consequence of conditional independence structure, and must derive its interpretation there. Furthermore, in Examples 2.22 and 2.23 we have seen distinct DAGs imposing exactly the same conditional independence structure.

This point is essential to understand if we are going to use observational data to infer the graphical structure of a Bayesian network model. To be sure, this is a useful and viable estimation problem, provided its limitations are understood. Fortunately, there is a very simple rule for determining when two DAGs impose the same conditional independent statements. Furthermore, this rule is a necessary and sufficient condition, which we now state.

Definition 2.10 Let G be a DAG. A *v-structure* is a subgraph consisting of three nodes, say a, b and c, such that a, b are parents of c, and there is no edge in G joining a and b (that is: $a \to c \leftarrow b$). The *skeleton* or *topology* of G is the undirected graph obtained by replacing all edges in G with undirected edges.

Two DAGs G and G' are *equivalent* if the following two conditions hold

(a) G and G' possess the same skeleton.

(b) G and G' possess the same v-structures.

The set of all DAGs which are equivalent to some DAG G forms an *equivalence class*. Any DAG G is equivalent to itself. Note that equivalent DAGs are necessarily defined on the same set of nodes V. ///

Example 2.24 We will apply Definition 2.10 to Examples 2.22 and 2.23.

In Figure 2.8, `DAG 1` and `DAG 2` are equivalent, since the skeletons are identical and neither DAG has a v-structure. In addition, no other DAG possesses the same skeleton, so `DAG 1` and `DAG 2` constitute a complete equivalence class (that is, there are no other DAGs equivalent to `DAG 1`).

In Figure 2.9, all four DAGs possess the same skeleton. Then `DAG 1`, `DAG 2` and `DAG 3` are equivalent, since none of these has a v-structure. On the other hand, `DAG 4` *does* have a v-structure, and so is not equivalent to the other three DAGs. There are no other DAGs which are equivalent to `DAG 1` (since only the DAGs shown in Figure 2.9 possess the same skeleton), so `DAG 1`, `DAG 2` and `DAG 3` constitute a complete equivalence class. □

The consequence of the equivalency of DAGs is quite profound.

Theorem 2.2 (Pearl *et al.* (1989)) Two DAGs G and G' defined on nodes V are equivalent (Definition 2.10) if and only if the following condition holds:

- Let g be any joint density on the nodes V which factors according to G. Then there exists a density g' on nodes V which factors according to G', and which satisfies $g = g'$.

///

Essentially, two equivalent DAGs impose the same set of conditional independence statements. Furthermore, suppose we use data to fit a density $\hat{g}$ which factors according to G, according to any optimal criteria. Then $\hat{g}$ will also factor according to any equivalent DAG G', meaning that we will have no basis on which to distinguish between G and G'.

2.5 Two Examples

In this section we present two extended examples illustrating the various concepts introduced in this chapter.

2.5.1 A simple gene regulatory network

Bayesian network models are often used to discern regulatory relationships in gene regulatory networks. Suppose observational data is used to fit a Bayesian network for 8 genes labeled $a, b, \ldots, g, h$, resulting in the DAG shown in Figure 2.10. We say a gene y is *downstream* from gene x if x is an ancestor of y (see Definition 2.1). In this case, x *regulates* y, possibly transitively. We will consider the following exercises.

(a) We first list all v-structures of the DAG, of which there are four:

$$b \to c \leftarrow d$$
$$b \to c \leftarrow h$$
$$d \to c \leftarrow h$$
$$g \to e \leftarrow c.$$

(b) Next, suppose a DAG is accepted as a true model of regulatory control. In this context, this means that all genes y which are downstream of any given gene x can be identified, assuming the inferred Bayesian network is correct. However, recall that observational data can only be used to infer an *equivalence class* of DAGs. This means that all DAGs in an equivalence class are equally compatible with the data. This being the case, any statement about regulatory order may be one of following three types:

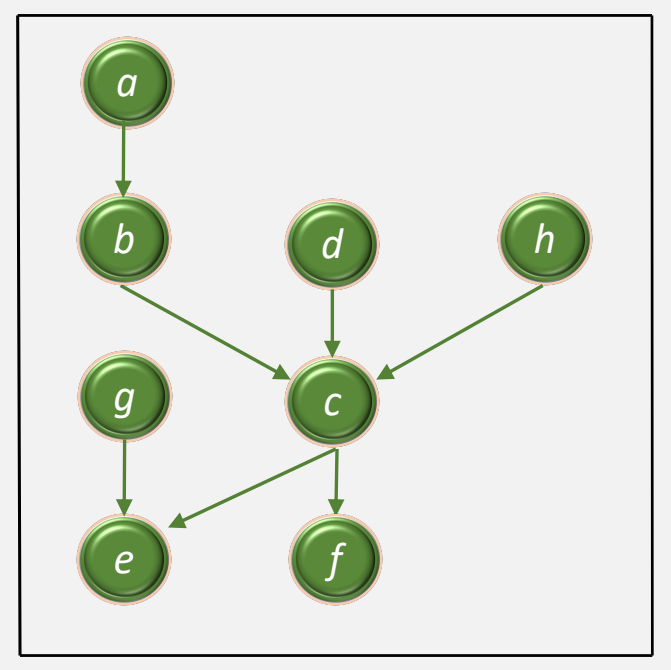

Figure 2.10: Sample DAG representing a gene regulatory network (Section 2.5.1).

Type A: Implied by the Bayesian network model (true of all equivalent DAGs).

Type B: Compatible with the Bayesian network model (true of some but not all equivalent DAGs).

Type C: Not compatible with the Bayesian network model (not true of any equivalent DAG).

Note that a DAG is equivalent to itself. As an exercise, we will determine the type (A, B or C) of each the following statements:

(i) c is downstream from h.
(ii) g is downstream from c.
(iii) h has no parents.
(iv) b has no parents.
(v) c has exactly three parents.
(vi) f is downstream from a.

Technically, to solve this type of problem it is important to understand how a DAG can be modified to produce a distinct but equivalent DAG, or to understand whether or not this operation is possible. First note that by Definition 2.10, two equivalent DAGs must have the same skeleton. This means that the direction of an edge can be changed,

but other than this, no edges can be added or removed. In addition, a switch in the direction of an edge cannot result in the removal or addition of a v-structure (otherwise, the DAG would not be equivalent, according to Definition 2.10). This is why it will be useful to identify all v-structures, as we have done in Part (a).

(i) In the original DAG, c is downstream from h, so the statement is not Type C. If there is an equivalent DAG in which c is *not* downstream of h, then the statement will be Type B. However, such a DAG could only be produced by reversing edge $h \to c$, which is part of a v-structure (two v-structures, actually). Since this operation would remove a v-structure, the resulting DAG will not be equivalent. Therefore, the statement is **Type A**.

(ii) g is not downstream from c in the original DAG, so the statement cannot be Type A. Furthermore, g can only be downstream of c if the edge $e \to g$ is reversed. However, this edge is part of the v-structure $g \to e \leftarrow c$, and so cannot be reversed to produce an equivalent DAG. Therefore the statement is **Type C**.

(iii) h has no parents in the original DAG. Furthermore, h shares an edge with node c only. However, the edge $h \to c$ is part of a v-structure, and cannot be reversed to produce an equivalent DAG. Therefore, the statement is **Type A**.

(iv) b has one parent, a, in the original DAG, so the statement cannot be Type A. Suppose edge $a \to b$ is reversed (so that now b has no parents). This edge is not part of a v-structure, and so none are removed. Furthermore, reversing edge $a \to b$ does not create any new v-structures, since a is not a child of another node. This means the equivalence class contains at least one DAG for which the statement is true, and at least one DAG for which the statement is false. Therefore the statement is **Type B**.

(v) All parents of c are part of v-structures pointing to c. Deletion or addition of any other parent would add or delete a v-structure. Since c has three parents in the original DAG, the statement is **Type A**.

(vi) f is downstream of a in the original DAG. In the discussion of statement (iv) above, it was argued that edge $a \to b$ could be reversed to produce an equivalent DAG. However, f would no longer be downstream from a in this DAG, so the statement is **Type B**.

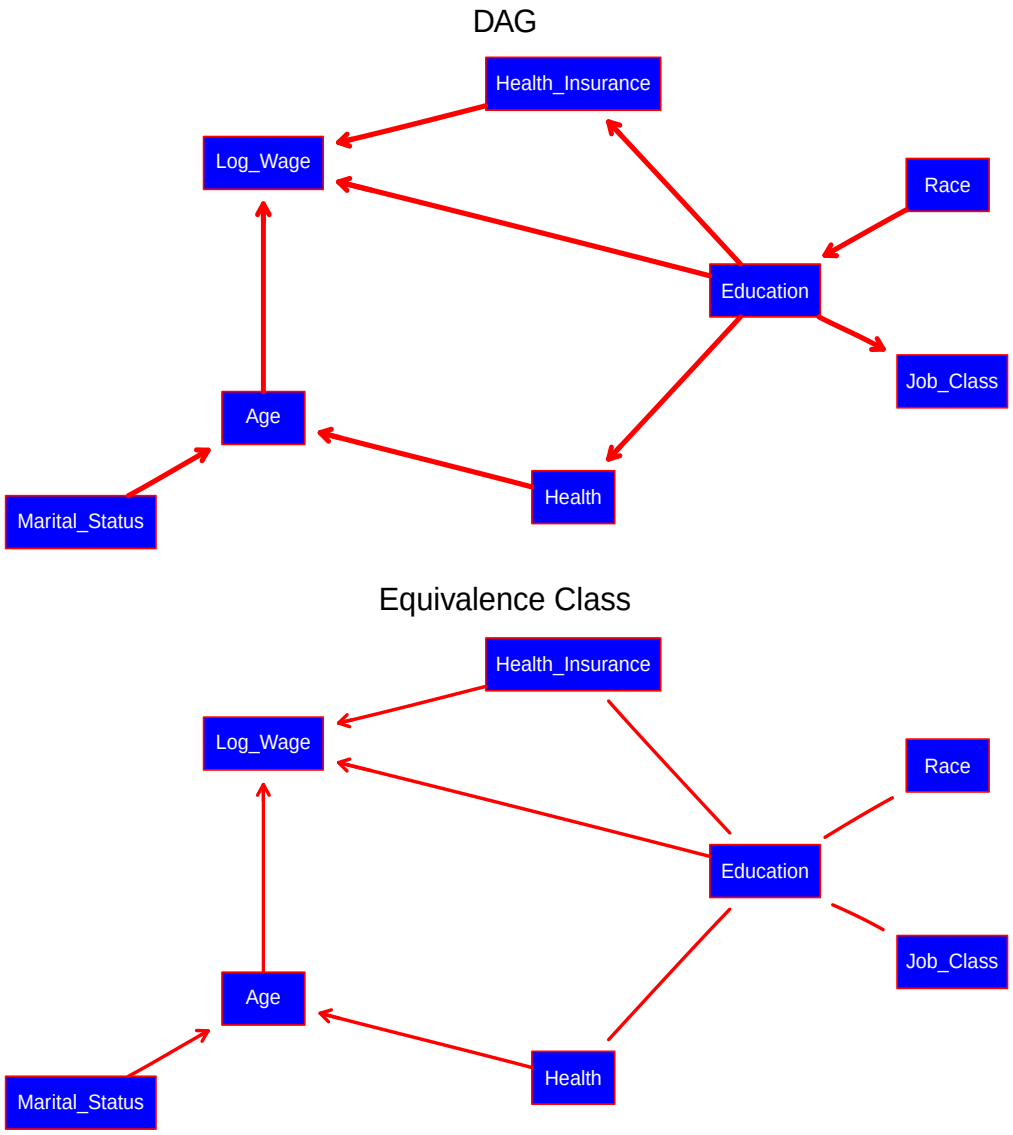

Figure 2.11: Bayesian network model fit with *Mid-Atlantic Wage Data* (Section 2.5.2). The original DAG is shown, as well as a schematic representation of the equivalence class

2.5.2 Mid-Atlantic wage data

We make use of the data set `Wage` included in the `ISLR` R-package (`https://cran.r-project.org/`). Subtitled *Mid-Atlantic Wage Data*, it contains wage and other data for 3000 male workers in the Mid-Atlantic region. See James *et al.* (2013) for more details. Eight variables from this data were used to fit a Bayesian network model, using the `hc(...)` function from

the `bnlearn` R-package (Scutari, 2010). We will discuss methods of fitting Bayesian networks in later chapters, but for now we will simply show the resulting DAG (Figure 2.11, top plot).

The log-transformed wage is given in the node labeled `Log_Wage`. The remaining nodes are of various types. `Age` is the worker's age in years. `Marital_Status` is a categorical variable with levels `Never Married`, `Married`, `Widowed`, `Divorced` and `Separated`. `Education` is a categorical variable with levels `< HS Grad`, `HS Grad`, `Some College`, `College Grad` and `Advanced Degree`. `Race` is a categorical variable with levels `White`, `Black`, `Asian` and `Other`. `Job_Class` is a categorical variable with levels `Industrial` and `Information`.

We note that Bayesian network models are flexible with regard to data type, and a single model often contains multiple types. This does not greatly affect their structure or interpretation.

In Section 2.5.1 it was made clear that the interpretation of a Bayesian network must take into account the entire equivalence class of a DAG. In Figure 2.11 a schematic representation of this equivalence class is shown in the bottom plot. This is constructed by replacing any edge of the original DAG with an undirected edge if there exists an equivalent DAG in which that edge is reversed. This is obtainable by converting any edge to an undirected edge if it is not part of a v-structure.

Thus, the undirected edges of the equivalence class representation are those edges which can be reversed in the original DAG to produce an equivalent DAG, following the technique used in Section 2.5.1. It must be stressed, however, that the choices of which edges to reverse cannot be made independently. For example, the original DAG contains edges

$$\texttt{Race} \rightarrow \texttt{Education} \quad \text{and} \quad \texttt{Education} \rightarrow \texttt{Job\_Class}.$$

Both of these edges are converted to undirected edges in the equivalent class representation, so each are represented in the equivalence class in both the original and reversed directions. However, suppose we reverse the edge $\texttt{Education} \rightarrow \texttt{Job\_Class}$. We will have then created a new v-structure:

$$\texttt{Race} \rightarrow \texttt{Education} \leftarrow \texttt{Job\_Class},$$

and the resulting DAG will not be equivalent to the original DAG. Of course, if we also reverse the edge $\texttt{Race} \rightarrow \texttt{Education}$ we now have path

$$\texttt{Race} \leftarrow \texttt{Education} \rightarrow \texttt{Job\_Class},$$

which is not a v-structure, and the resulting DAG will be equivalent to the original.

Interpreting Causality We now consider what the DAG tells us about the causal relationships among the nodes.

(a) First consider the node `Race`. From Figure 2.11 we can see it has child `Education`, no parents, and no other parents of its child. Its Markov blanket is therefore (by Definition 2.4):

$$B_{\texttt{Race}} = \{\texttt{Education}\}.$$

This means that conditional on `Education`, `Race` is independent of all remaining nodes. In particular, we have conditional independence statement:

$$(\texttt{Race} \perp\!\!\!\perp \texttt{Log\_Wage}) \mid \texttt{Education}.$$

In other words, `Log_Wage` depends on `Race`, but that dependence disappears once `Education` is taken into account. This means that wages are determined not by race but by education level. Thus, if there is a difference in wages between races, this is because there is a difference in education levels between races.

(b) We can reach a similar conclusion about the node `Job_Class`. It has one parent, `Education`, no children, and therefore no other parents of children. By Definition 2.4 the Markov blanket is therefore

$$B_{\texttt{Job\_Class}} = \{\texttt{Education}\},$$

and, as for `Race`, we have the conditional independence statement

$$(\texttt{Job\_Class} \perp\!\!\!\perp \texttt{Log\_Wage}) \mid \texttt{Education}.$$

(c) When we examine the DAG, it appears as though the node `Education` is very influential. It has child and parent sets

$$C_{\texttt{Education}} = \{\texttt{Health\_Insurance}, \texttt{Log\_Wage}, \texttt{Health}, \texttt{Job\_Class}\},$$
$$P_{\texttt{Education}} = \{\texttt{Race}\}.$$

In addition, the node `Log_Wage` is a child of `Education`, and has parents `Health_Insurance` and `Age`. While `Health_Insurance` is already included in $C_{\texttt{Education}}$, `Age` is included in neither $C_{\texttt{Education}}$ or $P_{\texttt{Education}}$, but by Definition 2.4 is included in the Markov blanket. This gives Markov blanket:

$$B_{\texttt{Education}}$$
$$= \{\texttt{Health\_Insurance}, \texttt{Log\_Wage}, \texttt{Health}, \texttt{Job\_Class}, \texttt{Race}, \texttt{Age}\},$$

which includes all nodes except for `Marital_Status` and `Education` itself. This suggests that `Education` is in some sense a highly influential node.

(d) The node `Marital_Status` has no parents; one child, `Age`; and one parent of a child, `Health`. The Markov blanket of `Marital_Status` is therefore

$$B_{\texttt{Marital\_Status}} = \{\texttt{Age}, \texttt{Health}\}.$$

This imposes the conditional independence statement

$$(\texttt{Marital\_Status} \perp\!\!\!\perp \texttt{Log\_Wage}) \mid \{\texttt{Age}, \texttt{Health}\}. \tag{2.20}$$

This has an interesting interpretation. It has been observed that higher wages tend to be positively associated with marriage. However, conditional independence statement (2.20) suggests that this is simply because married people tend to be older than single people, and wages almost universally increase with age.

Bibliography

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, **14**(3), 462–467.

Cowell, R. G., Dawid, P., Lauritzen, S. L., and Spiegelhalter, D. J. (2006). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer Science & Business Media.

Evans, A. S. (1978). Causation and disease: A chronological journey: The Thomas Parran Lecture. *American Journal of Epidemiology*, **108**(4), 249–258.

Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron*, **3**, 329–332.

Greenland, S., Pearl, J., Robins, J. M., *et al.* (1999). Causal diagrams for epidemiologic research. *Epidemiology*, **10**, 37–48.

Hogg, R. V., McKean, J. M., and Craig, A. T. (2018). *Introduction to Mathematical Statistics (What's New in Statistics)*. Pearson, 8th edition.

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). *Graphical Models with* `R`. Springer Science & Business Media.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling*. Guilford Publications.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.

Nagarajan, R., Scutari, M., and Lèbre, S. (2013). *Bayesian Networks in* `R`. Springer.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, 1985*, pages 329–334.

Pearl, J. (1986a). Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, **29**(3), 241–288.

Pearl, J. (1986b). On evidential reasoning in a hierarchy of hypotheses. *Artificial intelligence*, **28**(1), 9–15.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, California.

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

Pearl, J. (2012). The do-calculus revisited. *arXiv preprint arXiv:1210.4852*.

Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier.

Pearl, J. *et al.* (2009). Causal inference in statistics: An overview. *Statistics Surveys*, **3**, 96–146.

Pearl, J., Geiger, D., and Verma, T. (1989). Conditional independence and its representations. *Kybernetika*, **25**(7), 33–44.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley and Sons.

Ross, S. (2014). *Introduction to Probability Models*. Elsevier Science.

Ross, S. M. (1996). *Stochastic Processes*. John Wiley and Sons, New York, NY, 2nd edition.

Rubin, D. B. (1990). [On the application of probability theory to agricultural experiments. Essay on principles. Section 9.] Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, **5**(4), 472–480.

Sasco, A., Secretan, M., and Straif, K. (2004). Tobacco smoking and cancer: a brief review of recent epidemiological evidence. *Lung Cancer*, **45**, S3–S9.

Scutari, M. (2010). Learning Bayesian networks with the `bnlearn` `R` package. *Journal of Statistical Software*, **35**(i03).

Index