Causal Inference and Graphical Models

Anthony Almudevar

March 19, 2021

Table of Contents

Association and Causality The difference between *association* and *causality* is readily understood in an intuitive way, but can be difficult to define precisely. Two events A and B are associated if the occurrence of one is predictive of the occurrence of the other. This relationship is symmetric.

A causal relationship implies association, but claims more. The sequential nature of causality is readily apparent. If A causes B, we expect A to *precede* B. The street is wet because it rained. But a wet street does not cause rain. The latter hypothesis can be ruled out by the many observations of the sequence in which these two events occur.

Example 1 Of course, sequence does not by itself imply causality. Consider the following sequence of events:

A: It rains.
B: The street is wet.
C: A rainbow appears.

We can usually expect B to precede C, but B does not cause C. Their relationship is associative, but not causal. On the other hand, B and C are both caused by A.

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) Is observable through statistical inference.

Causality in the Service of Empirical Science As we will see, there are more than one meanings of causality which satisfy these requirements. So the approach taken here is not the development of foundational characterizations of causality, but rather the development of statistical techniques able to impose additional explanatory structure onto models of association, in the service of hypothesis driven investigation.

Experimental Observation of Causality The phrase "A causes B" 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). If we treat "A causes B" as a hypothesis, this might be resolvable using a suitably defined experiment.

Example 2 [Chemical Reaction - Version 1] To clarify this idea, consider the following scientific problem. 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

$$P(X_1 = X_2 = 1) \approx 0.75,$$
$$P(X_1 = X_2 = 0) \approx 0.25.$$

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, which occurred in approximately 75% of the trials.

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 data can detect association, but can not resolve a causal hypothesis. Suppose, however, that experimental techniques exists which can either force or suppress the presence of either agent, irrespective of other system variables. 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.

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

$$X_1 = -$$
$$X_1 = +$$
$$X_2 = -$$
$$X_2 = +.$$

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

$$P(X_2 = 1 \mid X_1 = -) \approx 0,$$
$$P(X_2 = 1 \mid X_1 = +) \approx 1.$$
$$P(X_1 = 1 \mid X_2 = -) \approx 0.75,$$
$$P(X_1 = 1 \mid X_2 = +) \approx 0.75.$$

What does this tell us? From the observation data 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 of 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 mere association.

DEFINITION 1 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 sampled replications 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 the preceding example cannot be resolved by observational data. And this follows from the theory of Bayesian networks. So it's worth asking what types of causal hypotheses, if any, can be resolved using observational data. In fact, the motivation of most of the theory introduced here is the resolution of this question.

OR:

HOW MUCH CAUSALITY CAN BE INFERRED WITH OBSERVATIONAL DATA? It might be said that what makes a careful study of the theory essential is the fact that observational data is able to partially, but rarely completely, resolve the causal structure of a process. And it might be the case that auxiliary information of some kind can be used to complete the causal model. Consider a second example.

Example 3 [Chemical Reaction - Version 2] We consider an experimental environment similar to that of the previous example, 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:

$$A \to B \to C$$
$$A \to C \to B$$
$$B \to A \to C$$
$$B \to C \to A$$
$$C \to A \to B$$
$$C \to B \to A$$

What a Bayesian Network Can Do The Bayesian network model is 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 hypothesis. In particular, a Bayesian network model could consistently identify the middle agent. Here, the term *consistent* is used precisely, in the sense that the probability that the middle agent is identified correctly approaches 1 as the sample size n approaches ∞.

What a Bayesian Network Can't Do Identifying the middle agent reduced 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 models.

EXAMPLE 4 Suppose a certain individual visits a hospital. It is conceivable that this affects the probability that the 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 variables:

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

For the moment, leave open the properties of $X_1, \ldots, X_6$. They may be categorical, binary (`TRUE/FALSE`), or quantitative. For example, X_1 could simply be `TRUE` if the individual visits the hospital. However, in a more refined model it might be a *quantitative variable* equal to the *length* of the visit, and X_6 might be a *probability* of missing work. As will be seen, 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 is the *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. See Figure 1.

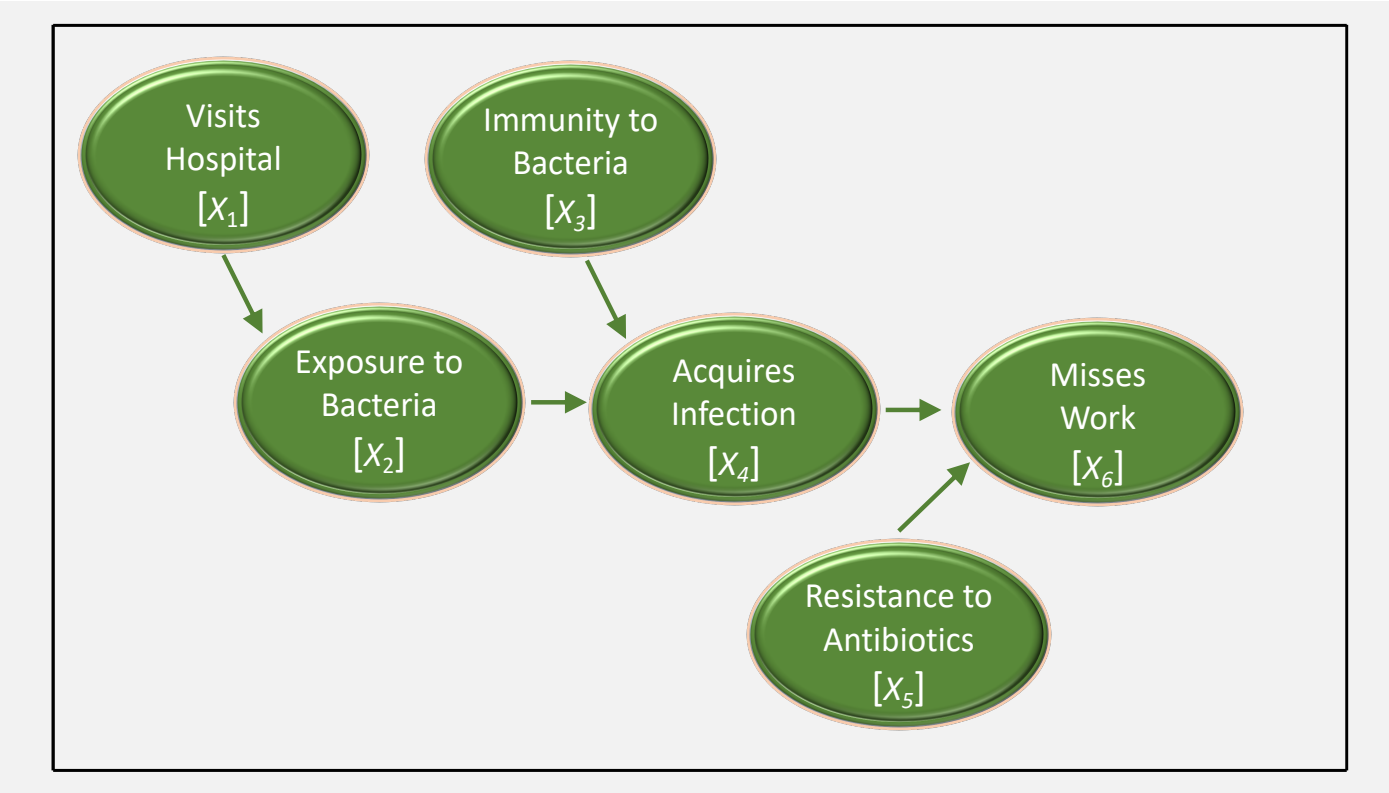


Figure 1: Introductory example.

DEFINITION 2 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 may be *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 both kinds of edges. The graph of Figure 1 is a *directed graph*.

We then introduce the following terminology:

- An directed edge points from a *parent* to a *child*.
- A directed graph contains *directed paths*, which are sequences of nodes in which consecutive nodes (left-to-right) form parent/child pairs. An example from Figure 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.
- A *directed path* joins an *ancestor* to a *descendent*.

- 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.
- Then let P_j be the set of random variables associated with the parents of node j. If node j is a founder, it has no parents. In this case we write $P_j = \{\} = \emptyset$. Similarly, let C_j be the set of random variables associated with the children of node j. If node j has no children we may write $C_j = \{\} = \emptyset$ ($\emptyset$ is the conventional symbol for an *empty set*).
- A directed path is a *cycle* if the first and last node are the same, with all other nodes appearing at most once. A *directed acyclic graph (DAG)* is a directed graph that contains no cycle. A DAG must contain at least one source and one sink. Figure 1 is a DAG.
- An *arc* is a sequence of nodes in which consecutive nodes are connected by edges of any kind.

The following is a partial list of the relationships given in Figure 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 *descendent* X_3.

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

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

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 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".

QUESTION If an arrow denotes causality, as it seems to do, and a visit to a hospital may cause one to miss work, why is there no arrow from 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.

A DAG represents not just a single sequential process, but several parallel sequential processes, which are only partially synchronized.

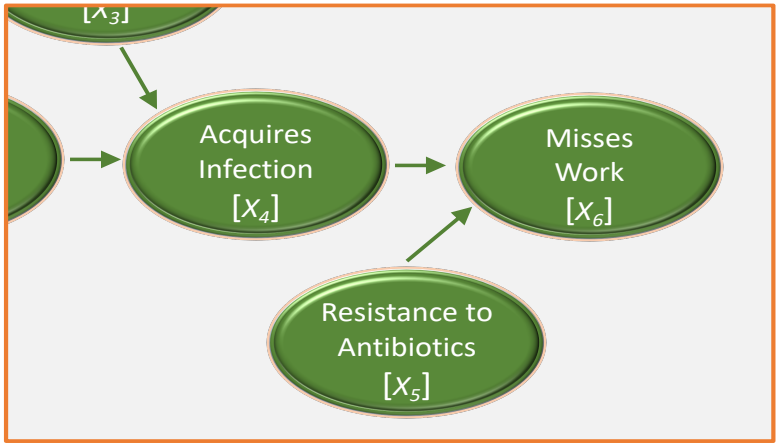


Figure 2: Section of Figure 1.

- We can, at least conceptually, assign a time T_i to each node X_i, representing the first time the variable label is observable. Even without knowing their exact values, they can, to some degree, be ordered.
- In particular, we can say 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 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 coherent. Such nodes, which represent states-of-nature, tend to appear in graphical models as founders.

It is useful to see a graph formed by a pedigree as an example of a DAG (Figure 3).

- The conventional terms *parent*, *child*, *ancestor* and *descendant* used for DAGs conform to their intuitive meanings with respect to a pedigree.

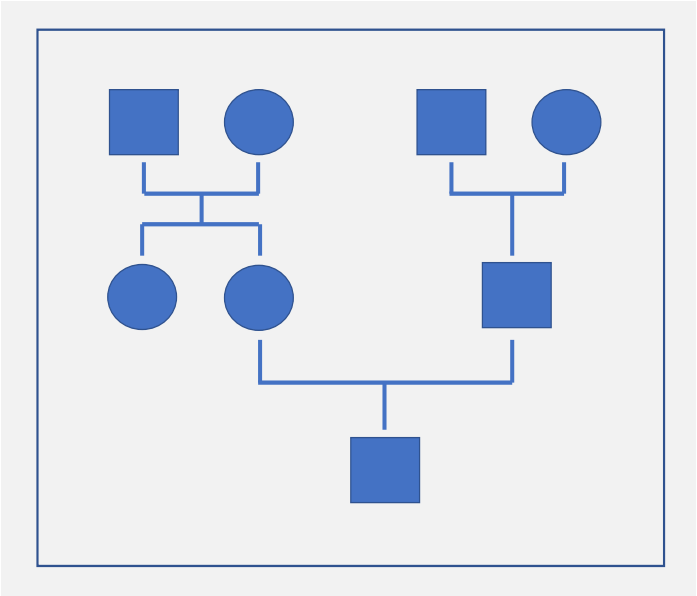


Figure 3: A pedigree graph is a DAG.

DEFINITION 3 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 5 For the DAG of Figure 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.

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

EXAMPLE 6 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.

The probabilistic structure of the Bayesian network can perhaps be made clear by comparison to a *Markov chain* (which, as will will see, is actually a special case 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*:

$$P(Z_{i+1} = a_{i+1} \mid Z_i = a_i, Z_{i-1} = a_{i-1}, \ldots, Z_1 = a_1) = P(Z_{i+1} = a_{i+1} \mid Z_i = a_i). \quad (1)$$

The process $Z_1, Z_2, Z_3, \ldots$ unfolds in time, so that Z_3 cannot be observed until Z_2 is observed, which cannot be observed until Z_1 is observed.

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. 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 (1).

The Memoryless Property

However, because of the memoryless property, expressed mathematically as Equation (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 $P(Z_n \mid Z_{n-1}) = P(Z_n \mid H_{n-1})$.

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 will be clarified in the next definition.

DEFINITION 4 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 variable 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 $F_{X,Y|Z}$, $F_{X|Z}$, $F_{Y|Z}$ and the joint and marginal cumulative distribution functions (CDF) of X, Y conditional on $Z = z$.

Examples of Conditional Independence

EXAMPLE 7 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 and not independent, unless we condition on N.

EXAMPLE 8 A dice is tossed independently three times. Let S_1, S_2, S_3 be the cumulative totals. Clearly, S_1 and S_3 are not independent. For example, the reader may verify the counter-example:

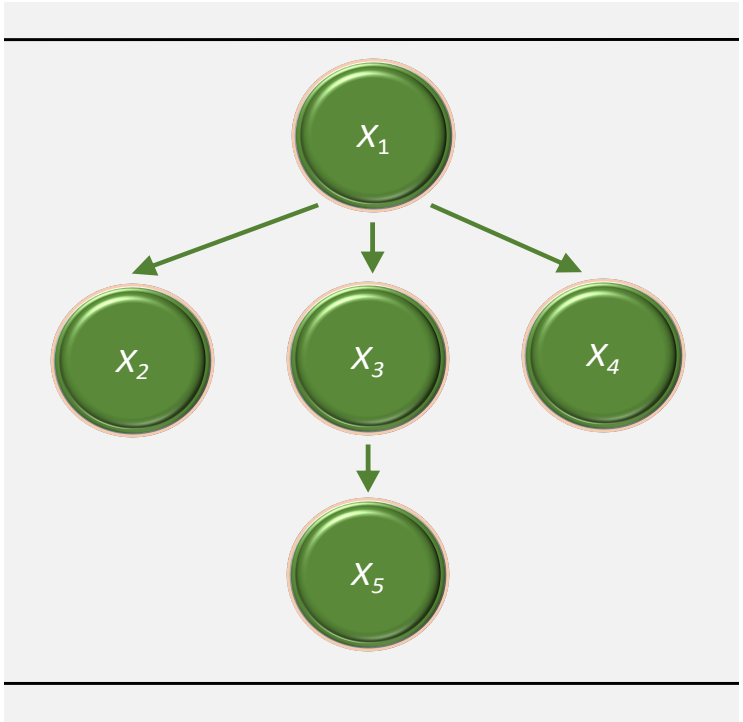
$$\begin{aligned} P(S_3 = 18 \mid S_1 = 6) &= 1/36, \text{ but} \\ P(S_3 = 18 \mid S_1 = 5) &= 0. \end{aligned}$$

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.

What exactly is a Bayesian Network? Despite the term *graphical model*, the *graph* need not be the most important object defining the Bayesian network (BN) model. It is sometimes helpful to think of a BN as nothing more than a type of joint distribution $g = g(x_1, x_2, \ldots, x_n)$ of random variables $X_1, X_2, \ldots, X_n$.

Then what role is played by the graph? A joint distribution g does not define a BN unless it satisfies certain constraints. Those constraints are imposed by the graph (the DAG, to be precise). Moreover, these constraints take the form of conditional independency statements. A DAG, therefore may be equivalently thought of as a list of conditional independency statements.

EXAMPLE 9 The following DAG (Figure 4) imposes the indicated list of conditional independence statements. Note that conditional independence statements can apply to sets of variables, grouped by brackets, for example, $\{X_2, X_4, X_5\}$.



$$\Longrightarrow$$

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

Figure 4: Example of DAG. Conditional independence statements generated by the DAG are shown to the right.

So far, are discussion has been purely descriptive, and the following questions come to mind.

QUESTION According to what rules does the DAG in Figure 4 generate the associated set of conditional independence statements?

QUESTION How can we construct a joint distribution which is constrained to conform to a specific set of conditional independence statements?

There is a quite deep theory able to resolve these questions, which will be discussed in due time. We will first, however, continue with this example. It is important to show that while the underlying mathematics can be quite formidable, the models themselves are usually quite intuitive.

We now consider how conditional independence statements are imposed by a DAG. The key is in the following definition.

DEFINITION 5 Given a DAG, 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, excluding node j.

Denote this set of nodes B_j. This set may also refer to the collection of random variables associated with the nodes in B_j, as necessitated by the context. (Note that this is the definition of a Markov blanket for DAGs. This definition may differ for other types of graphs.)

The rules for generating the conditional independence statements for the DAG of Figure 4 can now be stated. Let V be the set of all nodes. The conditional independence statements are then

$$\{X_j\} \perp\!\!\!\perp V - \{X_j\} - B_j \mid B_j, \;\; j = 1, \ldots, n$$

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

EXAMPLE 10 Consider node X_3 of Figure 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

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

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*.

DEFINITION 6 Suppose a and b are distinct nodes of a DAG. Let L be an arc between a and b. Then a node c 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 L is *blocked* by C if either

Rule 1: L contains a node in C that is not a collider; or

Rule 2: L contains a node z that is a collider, such that neither z nor any of its descendants is in C.

Then let A, B, C be three disjoint subsets of nodes. We say C *d-separates* A and B if any arc from any node $a \in A$ and $b \in B$ is *blocked* by C (equivalently, A and B are blocked by C).

The rule for generative conditional independence statements via d-separation is now simply:

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

EXAMPLE 11 Consider the DAG of Figure 1.

Case 1: 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 1 of Definition 6 is satisfied, so that A and B are blocked by C, which imposes conditional independence statement

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

Case 2: 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 from a node in A to a node in B passes through X_6, which will be a collider. Then note that in order for C to d-separate A and B, at least one of the two rules of Definition 6 must hold. Rule 1 cannot hold, since there are no nodes in C. On the other hand, all arcs contain a collider z such that neither z, nor any of its descendants, is in C. This is because there is a collider on all paths, but no nodes belong to C. We therefore conclude

$$\{X_5\} \perp\!\!\!\perp \{X_1, X_2, X_3, X_4\} \mid \{\} \text{ or } \{X_5\} \perp\!\!\!\perp \{X_1, X_2, X_3, X_4\}$$

which means that the nodes in A are independent of the nodes in B. As a technical detail, we note that although node X_4 has in-degree 2, and would be a collider on *some* arc, it is not a collider on any of the arcs joining nodes from A and B.

EXAMPLE 12 Consider the DAG of Figure 5.

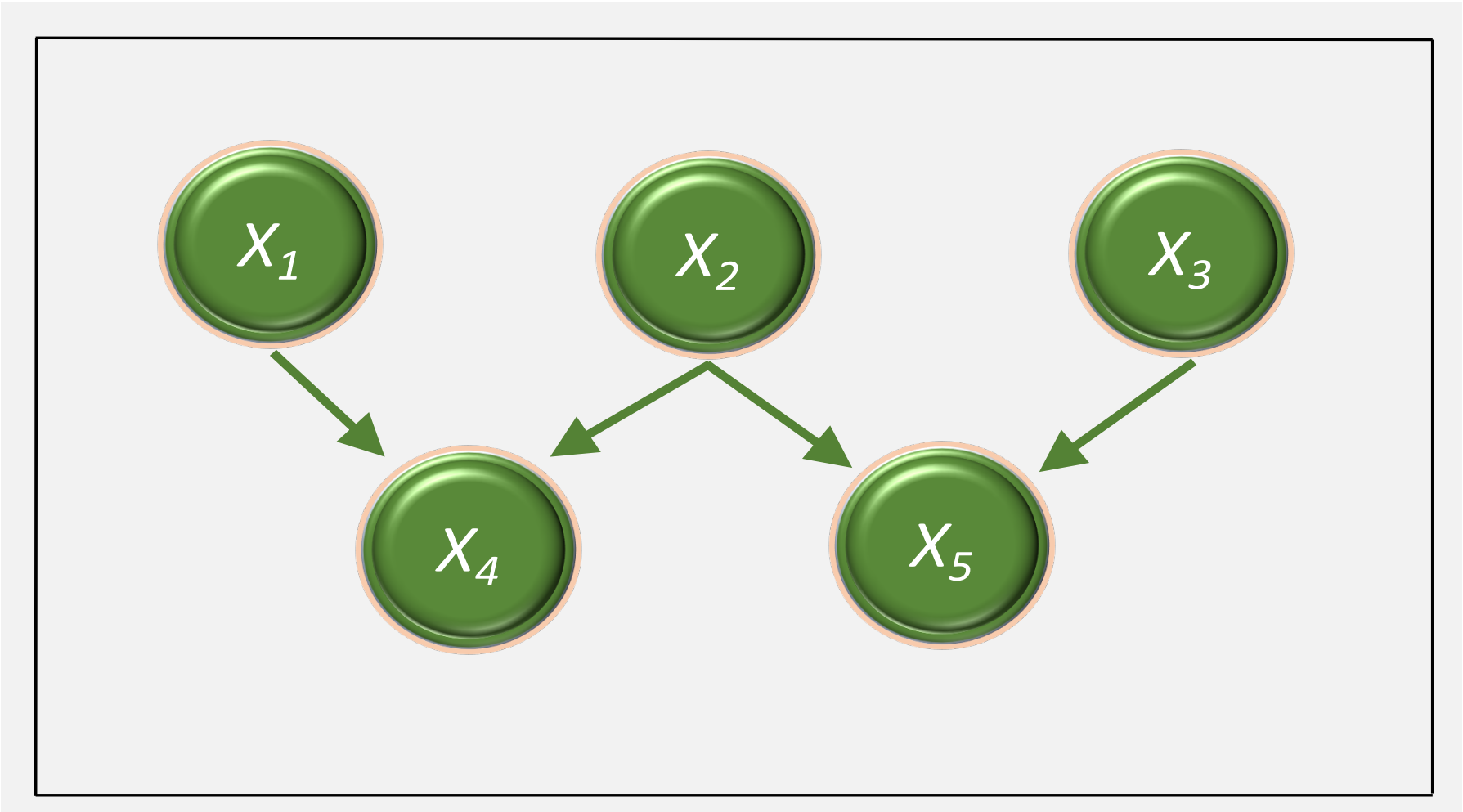


Figure 5: DAG used in Example 12.

Case 1: 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 5 the Markov blanket of X_1 is therefore $B_1 = \{X_2, X_4\}$. This imposes conditional independence statement

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

Case 2: 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

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

Case 3: Then set $C = \{X_4\}$. Does C d-separate $A = \{X_1\}$ and $B = \{X_3\}$? 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 2 of Definition 6 is satisfied, and we conclude that C d-separates A and B. The DAG therefore imposes conditional independence statement

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

Case 4: Next, set $C = \{X_5\}$. Does C d-separate $A = \{X_1\}$ and $B = \{X_3\}$? 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

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

Case 5: Next, set $C = \{X_4, X_5\}$. Does C d-separate $A = \{X_1\}$ and $B = \{X_3\}$? There are no non-colliders on the arc joining X_1 and X_3 which are in C, so Rule 1 is violated. Also, any collider on the arc is in C, so that Rule 2 is violated. Thus, we conclude that C does *not* d-separate A and B. We will discuss this case further in Example 15 below.

Case 6: Set $C = \{X_2, X_4, X_5\}$. Does C d-separate $A = \{X_1\}$ and $B = \{X_3\}$? There is only one arc joining X_1 and X_3. Rule 1 is satisfied contains a non-collider X_2 which is in C, and we conclude that C d-separates A and B. The DAG therefore imposes conditional independence statement

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

Case 7: Set $C = \{\}$. Does C d-separate A and B? The single arc joining X_1 and X_3 contains a node which is a collider z such that neither z, nor any of its descendants, is in C. Thus, Rule 2 holds, and we conclude that C d-separates A and B. The DAG therefore imposes conditional independence statement

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

So far, we have seen that a DAG implies a collection of conditional independence statements. The next problem to develop a method of constructing joint densities for the nodes which conform to those statements.

We will first use our intuition to build a model which conforms to the DAG in Figure 4, and to the associated set of conditional independence statements. To do this we will attempt to mimic a Markov chain. Let $\epsilon_1, \ldots, \epsilon_5$ be five independent random variables, of any kind.

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. The rule is simple. Each node inherits the values of its parents, 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 to Figure 4, 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?

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 devise to denote this, putting in square brackets any random variable or expression on which we are conditioning, and which we therefore which 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 $[\epsilon_1]$.

EXAMPLE 13 Consider the conditional independence statement:

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

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 holds.

EXAMPLE 14 Consider the conditional independence statement:

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

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 holds.

EXAMPLE 15 We can use the approach of Examples 13 and 14 to understand why, in Example 12 (Figure 5) 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 (Cases 3 and 5). Using the two rules of Examples 13-14 would give here

$$\begin{aligned} X_4 &= X_1 + X_2 + \epsilon_4, \\ X_5 &= X_2 + X_3 + \epsilon_5, \end{aligned} \tag{2}$$

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:

$$(X_1 + \epsilon_4) - (X_3 + \epsilon_5) = s - t. \tag{3}$$

How can we interpret Equation (3)? We can take $s - t$ to be constant, and then interpret (3) 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:

$$X_1 - X_3 \approx s - t. \tag{4}$$

It is easily verified that two independent random variables are no longer independent when conditioned on a constraint such as Equation (4).

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:

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

A simpler example makes the same point. If U_1, U_2 are independent random variables, and $U_1 + U_2 = U_3$, Then U_1, U_2 will not be independent conditional on $U_3 = u$. Figure 5 provides a more complex version of this effect.

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 expression simplifies considerable 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

$$g(x_1, x_{n+2}, \ldots, x_n) = \left(\prod_{j=1}^{n-1} g(x_{n-j+1} \mid x_{n-j}) \right) \times g(x_1). \tag{5}$$

The BN is 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 16 For the DAG in Figure 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 17 Founders are an important special case. Suppose X_1 is a founder. Then

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

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 BN, representable by the DAG shown in Figure 6.

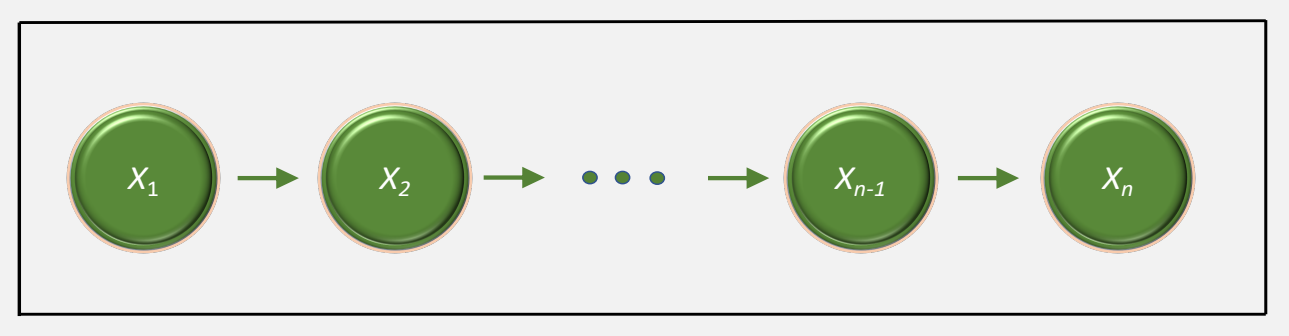


Figure 6: DAG representation of a Markov Chain.

Next, note that Equation (5), which gives the joint distribution of a Markov chain can be rewritten using the parent sets:

$$g(x_1, x_{n+2}, \ldots, x_n) = \prod_{j=1}^{n} g(x_j \mid P_j), \tag{6}$$

noting that the term $g(x_1)$ appearing in Equation (5) is represented in Equation (6) 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 6.

Theoretical basis for the Bayesian network model

At this point, we have enough to formally define a Bayesian network supported by a rigorous mathematical foundation.

FACTORIZATION THEOREM Let $V = (X_1, \ldots, X_n)$ be n random variables which label n nodes of a DAG, say $G = (V, E)$. Suppose the joint distribution can be factored in the following way:

$$g(x_1, x_{n+2}, \ldots, x_n) = \prod_{j=1}^{n} g(x_j \mid P_j), \tag{7}$$

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

Then all conditional independence statements of the form

$$A \perp\!\!\!\perp B \mid C$$

hold whenever C d-separates A and B.

In addition, all conditional independence statements of the form

$$\{X_j\} \perp\!\!\!\perp V - \{X_j\} - B_j \mid B_j, \;\; j = 1, \ldots, n$$

hold, where B_j is the Markov blanket of node X_j.

This formally defines the Bayesian network model.

Recall that in Example 3 there was interested 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 would be able to reduce the number of hypotheses from 6 to 2 (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 two hypotheses:

$$A \to C \to B$$
$$B \to C \to A.$$

In Example 2, a *cause* A was something that was either necessary or sufficient for the occurrence of *effect* B. In Example 3, we are really considering *conditional independence*, which is a concept that is different from, but related to, direct causality. Suppose the correct hypothesis of Example 3 is

$$A \to C \to B.$$

We can recognize this as defining a Bayesian network model, provided the construction specified by the Factorization Theorem holds. Furthermore, B has parent C, no children, and shares no children with another node. The Markov blanket of B is therefore $\{C\}$, and the following conditional independence statement holds:

$$A \perp\!\!\!\perp B \mid C$$

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 cause of B.

So far, we have seen

(1) A method of determining conditional independence statements imposed by a DAG;

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

Clearly, there is a very important third step. It almost goes without saying that the motivation for using graphical models is 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. 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 7 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 $P(d(\hat{\theta}_n, \theta^*) < \epsilon) \to 0$ for all $\epsilon > 0$ and $\theta^* \in \Theta$. We then say such an estimator is *consistent*.

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, but also additional parameters (which we call *auxilliary parameters*) defining the conditional distributions used in the factorized distribution of Equation (7).

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

(1) Conditional independence statements can be consistently tested;

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

So, the inference of a Bayesian network model can be decomposed into subproblems of one of two types:

(1) Estimation of conditional independence statements;

(2) Estimation of auxilliary parameters.

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

ANSWER The density of the Bayesian network given in Equation (7) 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.

To explore this issue, it helps to start with the simplest Bayesian network model that still retains some interesting structure.

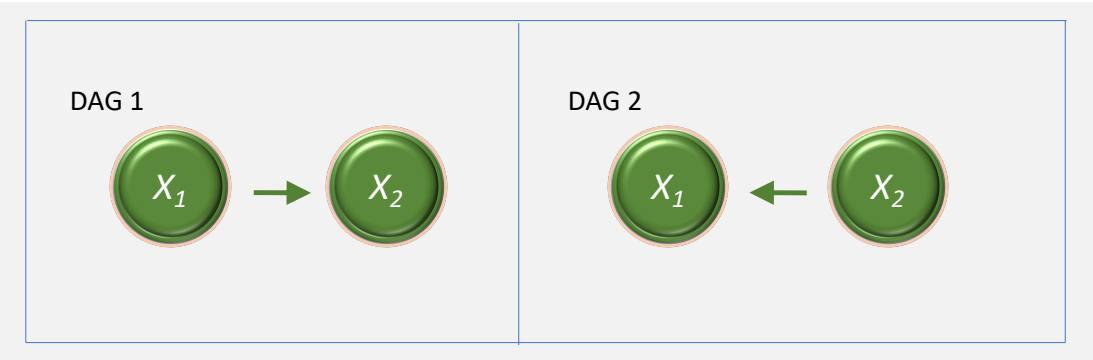


Figure 7: Examples of DAGs (two nodes and one edge).

If we apply the Factorization Theorem, the joint density for (X_1, X_2) imposed by `DAG 1` of Figure 7 would be, using Equation (7),

$$\begin{aligned} g(x_1, x_2) &= g(x_1 \mid P_1)g(x2 \mid P_2) \\ &= g(x_1 \mid \{\})g(x_2 \mid x_1) \\ &= g(x_1)g(x_2 \mid x_1) \\ &= g(x_1, x_2). \end{aligned}$$

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 7, we similarly have

$$\begin{aligned} g(x_1, x_2) &= g(x_1 \mid P_1)g(x2 \mid P_2) \\ &= g(x_1 \mid x_2)g(x_2 \mid \{\}) \\ &= g(x_1 \mid x_2)g(x_2) \\ &= g(x_1, x_2). \end{aligned}$$

The structure is the same as for `DAG 1`. Either model admits any form of dependence between X_1 and X_2, and are 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.

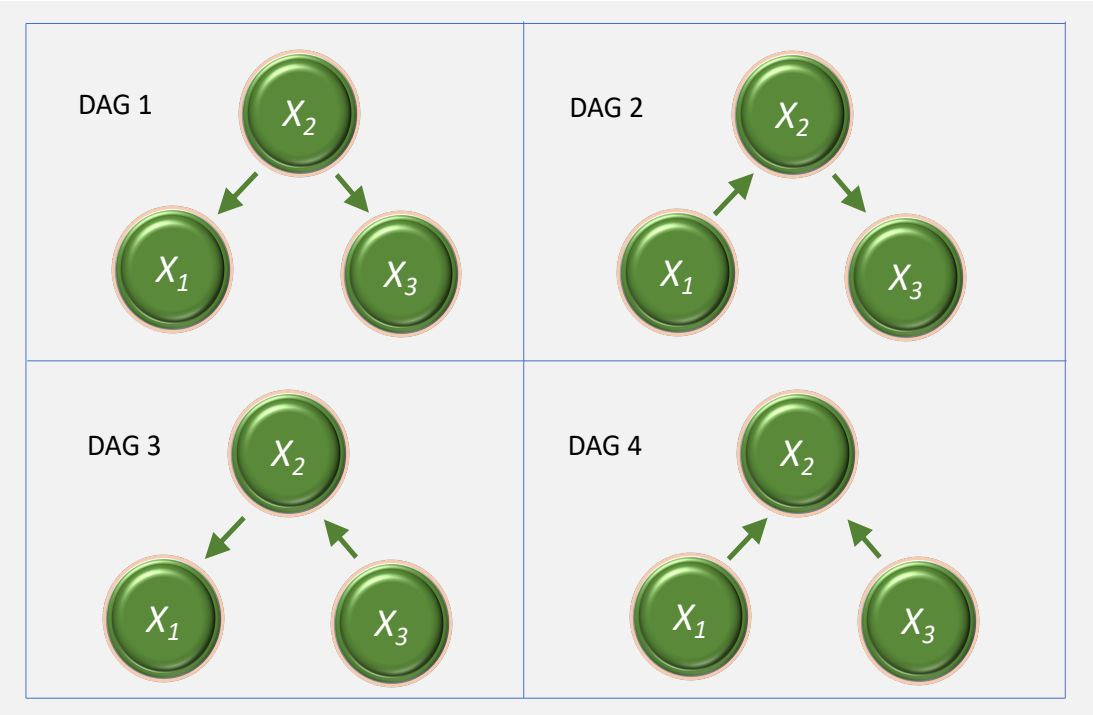


Figure 8: Examples of DAGs (three nodes and two edges).

If we apply the Factorization Theorem to `DAG 1` of Figure 8 we have a joint distribution for (X_1, X_2, X_3) of the form

$$\begin{aligned} g(x_1, x_2, x_3) &= g(x_1 \mid P_1)g(x2 \mid P_2)g(x3 \mid P_3) \\ &= g(x_1 \mid x_2)g(x_2 \mid \{\})g(x_3 \mid x_2) \\ &= g(x_1 \mid x_2)g(x_2)g(x_3 \mid x_2). \end{aligned}$$

If we divide this expression by $g(x_2)$ we may write equivalently:

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

It is not hard to verify that when this expression is compared to Definition 4 (of conditional independence) we may claim $X_1 \perp\!\!\!\perp X_3 \mid X_2$.

We can reach the same conclusion using Markov blankets. For `DAG 1` of Figure 8 we have $P_1 = \{X_2\}$, $C_1 = \{\}$, and node X_1 shares no children with other parents (Definition 5). The Markov blanket for node X_1 is therefore $B_1 = \{X_2\}$, and by the Factorization Theorem the conditional independence statement $X_1 \perp\!\!\!\perp X_3 \mid X_2$, which we also were able to conclude by direct construction of the distribution function of (X_1, X_2, X_3).

The structure of `DAG 2` and `DAG 3` of Figure 8 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` (Figure 8). If we construct the joint distribution for (X_1, X_2, X_3) using the Factorization Theorem (Equation (7)) we obtain the form

$$\begin{aligned} g(x_1, x_2, x_3) &= g(x_1 \mid P_1)g(x2 \mid P_2)g(x3 \mid P_3) \\ &= g(x_1 \mid \{\})g(x_2 \mid x_1, x_3)g(x_3 \mid \{\}) \\ &= g(x_1)g(x_2 \mid x_1, x_3)g(x_3). \end{aligned}$$

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

This point is essential to understand if we are going to use observation 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 8 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) DAGs G and G' possess the same skeleton.

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

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

The consequence of equivalency of DAGs is quite profound.

THEOREM [VERMA & PEARL, 1990] Two DAGs G and G' defined on nodes V are equivalent if and only if the following condition holds:

(a) 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'.

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 9. We say a gene y is *downstream* from gene x if x is an ancestor of y. In this case, x *regulates* y, possibly transitively. We will consider the following exercises.

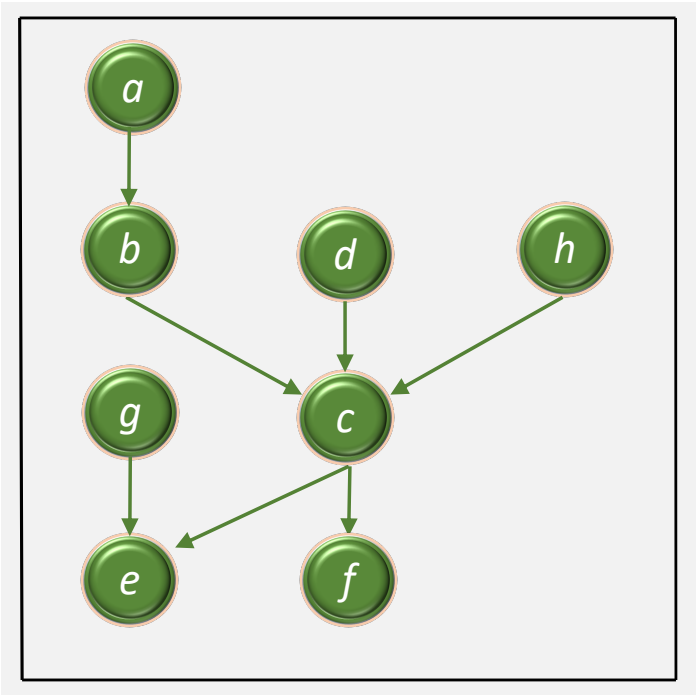


Figure 9: Sample DAG representing a gene regulatory network.

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

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

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:

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 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). This is why it will be useful to identify all v-structures, as we have done above.

(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**.

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 (James, G. *et al. Introduction to Statistical Learning*, Springer).

Eight variables from this data were used to fit a Bayesian network model, using the `hc(...)` function from the `bnlearn` R-package (Scutari, M. (2010) *Learning Bayesian networks with the* `bnlearn` `R` *package*, Journal of Statistical Software, 35(i03)) We will discuss methods of fitting Bayesian networks in later chapters, but for now we will simply show the resulting DAG (Figure 10, top plot).

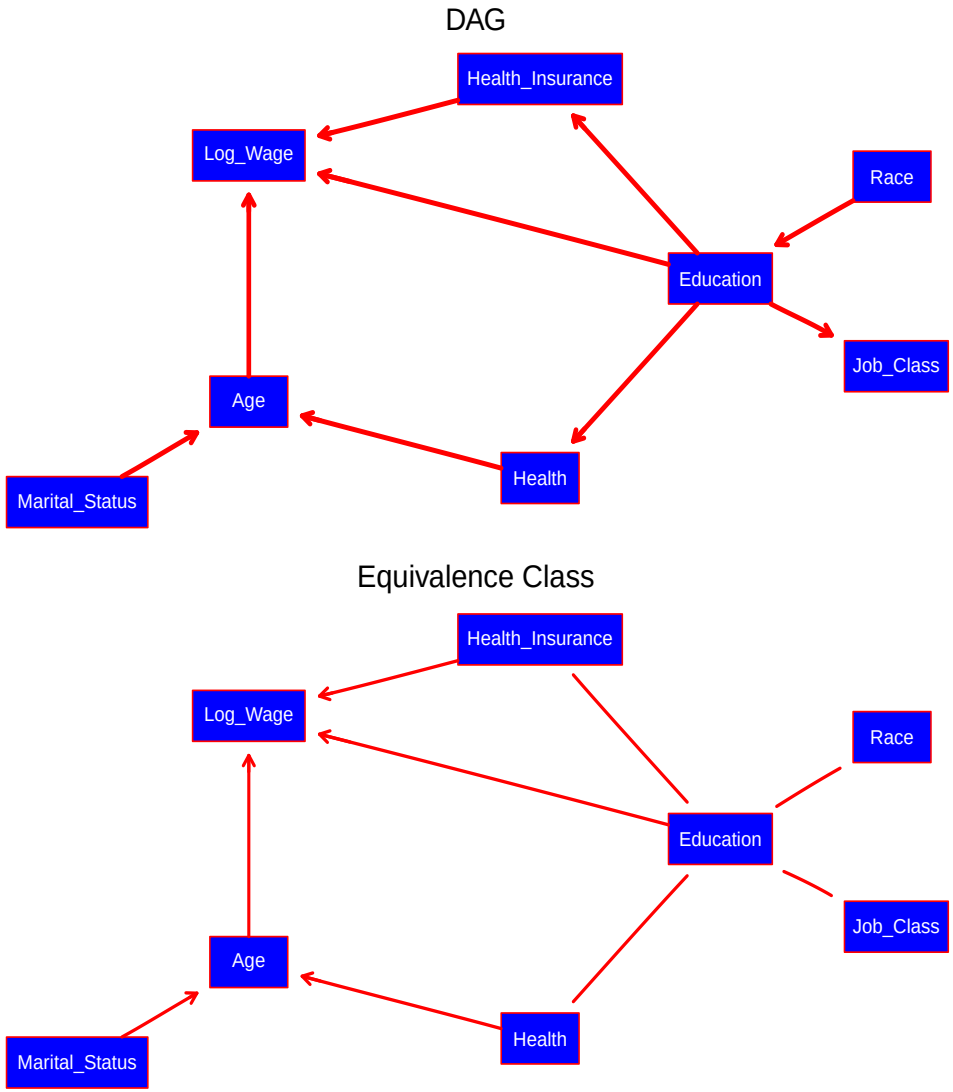


Figure 10: Bayesian network model fit with *Mid-Atlantic Wage Data*. The original DAG is shown, as well as a schematic representation of the equivalence class

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.

We have already seen that the interpretation of a Bayesian network must take into account the entire equivalence class of a DAG. In Figure 10 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 the prevous example. It must be stressed, however, that the choices of which edges to reverse cannot be made independently. For example, the original DAG contains edges

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

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 $\texttt{Education} \to \texttt{Job\_Class}$. We will have then created a new v-structure:

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

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

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

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 10 we can see it has child `Education`, no parents, and no other parents of its child. Its Markov blanket is therefore:

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

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.

(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. 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 is included in the Markov blanket. This gives Markov blanket:

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

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

$$(\texttt{Marital\_Status} \perp\!\!\!\perp \texttt{Log\_Wage}) \mid \{\texttt{Age}, \texttt{Health}\}. \tag{8}$$

This has an interesting interpretation. It has been observed that higher wages tend to be positively associated with marriage. However, conditional independence statement (8) suggests that this is simply

because married people tend to be older than single people, and wages almost universally increase with age.