Learning Bayesian Networks

Introduction to Bayesian Networks

Joint Probability Distributions

Definition 1.1 Suppose we have sample space Ω containing n distinct elements. That is:

$$\Omega = \{e_1, e_2, \dots, e_n\}.$$

A function which assigns a real number $P(E)$ to each event $E \subseteq \Omega$ is called probability function on the set of subsets of Ω if it satisfies the following conditions:

1. $0 \leq P(\{e_i\}) \leq 1 \quad \text{for} \quad 1 \leq i \leq n$
2. $P(\{e_1\}) + P(\{e_2\}) + \cdots + P(\{e_n\}) = 1$
3. For each event $E = \{e_{i_1}, e_{i_2}, \dots, e_{i_k}\}$ that is not an elementary event,
   $P(E) = P(\{e_{i_1}\}) + P(\{e_{i_2}\}) + \cdots + P(\{e_{i_k}\})$

The pair (Ω, P) will be denoted as ***probability space***.

Theorem 1.1 Let (Ω, P) be a probability space. Then

1. $P(\Omega) = 1$.
2. $0 \leq P(E) \leq 1$ for every $E \subseteq \Omega$.
3. For E and $F \subseteq \Omega$ such that $E \cap F = \emptyset$,
   $P(E \cup F) = P(E) + P(F)$.

Definition 1.2 Let E and F be events such that $P(F) \neq 0$. Then the conditional probability of E given F, denoted $P(E|F)$, is given by:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Definition 1.3 Two events are independent if one of the following hold:

1. $P(E|F) = P(E)$ and $P(E) \neq 0, P(F) \neq 0$
2. $P(E) = 0$ or $P(F) = 0$

Definition 1.4 Two events E and F are conditionally independent given G if $P(G) \neq 0$ and one of the following holds:

1. $P(E|F \cap G) = P(E|G)$ and $P(E|G) \neq 0, P(F|G) \neq 0$
2. $P(E|G) = 0$ or $P(F|G) = 0$

Definition 1.8 Let a set of n random variables $V = \{X_1, X_2, \dots, X_n\}$ be specified such that each X_i has a countably infinite space. A function, that assigns a real number $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$ to every combination of values of the x_i's such that the value of x_i is chosen from the space of X_i, is called joint probability distribution of the random variables in V if it satisfies the following conditions:

1. For every combination of values of the x_i's:
   $0 \leq P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \leq 1$
2. We have:
   $\sum_{x_1, x_2, \dots, x_n} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = 1$

Suppose we have n events $E_1, E_2, \ldots, E_n$ such that $E_i \cap E_j = \emptyset$ for $i \neq j$ and $E_1 \cup E_2 \cup \ldots \cup E_n = \Omega$. Such events are called ***mutually exclusive and exhaustive***.

The law of total probability says that for any event $F \subset \Omega$ we have:
$P(F) = \sum_{i=1}^{n} P(F|E_i)P(E_i)$

Theorem 1.2 (Bayes) Given two events E and F such that $P(E) \neq 0$ and $P(F) \neq 0$ we have:
$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$
Furthermore, given n mutually exclusive and exhaustive events E_1, E_2, …, E_n such that $P(E_i) \neq 0$ for all i, we have for $1 \leq i \leq n$:
$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{\sum_{j=1}^{n} P(F|E_j)P(E_j)}$

Definition 1.5 Given a probability space (Ω, P), a ***random variable*** X is a function of Ω. The set of values X can assume is called ***the space*** of X. A random variable is said to be ***discrete*** if its space is finite or countable.

Theorem 1.3 Let a set of random variables V be given and let a joint probability distribution of the variables in V be specified according to Definition 1.8. Let Ω be the Cartesian product of the sets of all possible values of the random variables. Assign probabilities to elementary events in Ω as follows:
$\hat{P}(\{(x_1, x_2, \ldots, x_n)\}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$
These assignments result in a probability function on Ω according to ***Definition 1.1***. Furthermore, if we let $\hat{X}_i$ denote a function (random variable in the classical sense) on this sample space which maps each tuple in Ω to the value of x_i in that tuple, then the joint probability distribution of the $\hat{X}_i$'s is the same as the originally specified joint probability distribution.

Markov Condition

Definition: ***ancestral ordering*** is such ordering for which if Y is descendant of X then Y is on the right of X

Definition 1.9 Suppose we have a joint probability distribution P of the random variables in some set V and a DAG $\mathbb{G} = (V, E)$. We say that $(\mathbb{G}, P)$ satisfies ***the Markov condition*** if for each variable $X \in V$, $\{X\}$ is conditionally independent of the set of all of its non-descendants given the set of all its parents. If we denote the sets of parents and non-descendants of X by PA_X and ND_X, respectively, then

$$I_P(\{X\}, ND_X|PA_X)$$

When $(\mathbb{G}, P)$ satisfies *the Markov condition*, we say $\mathbb{G}$ and P satisfy Markov condition with each other. If X is a root, then its parent set PA_X is empty. So in this case it means that the Markov condition means $\{X\}$ is independent of ND_X - $I_P(\{X\}, ND_X)$. But $I_P(\{X\}, ND_X|PA_X)$ implies $I_P(\{X\}, B|PA_X)\ \forall\ B \subseteq ND_X$. We have $PA_X \subseteq ND_X$. So we can rewrite the Markov condition as:

$$I_P(\{X\}, ND_X - PA_X|PA_X)$$

Theorem 1.4 If $(\mathbb{G}, P)$ satisfies *the Markov condition*, then P is equal to the product of its conditional distributions of all nodes given values of their parents, whenever these conditional distributions exist.

Proof: We prove the case where P is discrete. Order the nodes in their ancestral ordering. Let $X_1, X_2, \dots, X_n$ be the resultant ordering. For a given set of values $x_1, x_2, \dots, x_n$ let $\mathbf{pa}_i$ be the subset of these values containing the values of X_i's parents. We need to show that whenever $P(\mathbf{pa}_i) \neq 0$ for $1 \leq i \leq n$,

$$P(x_n, x_{n-1}, \dots, x_1) = P(x_n|\mathbf{pa}_n)P(x_{n-1}|\mathbf{pa}_{n-1}) \dots P(x_1|\mathbf{pa}_1)$$

We show this using induction on the variables of the network. Assume for some combination of values of the $x_i's$, that $P(\mathbf{pa}_i) \neq 0$ for $1 \leq i \leq n$.

Induction base $(n = 1)$:
Since PA_1 is empty $P(x_1) = P(x_1|\mathbf{pa}_1)$.

Induction hypothesis $(n = i)$:
Suppose for this combination of values of the $x_i's$ that:
$P(x_i, x_{i-1}, \dots, x_1) = P(x_i|\mathbf{pa}_i)P(x_{i-1}|\mathbf{pa}_{i-1}) \dots P(x_1|\mathbf{pa}_1)$

Induction Step: Prove $n = i + 1$ assuming that the hypothesis for $n = i$ is true
$P(x_{i+1}, x_i, \dots, x_1) = P(x_{i+1}|\mathbf{pa}_{i+1})P(x_i|\mathbf{pa}_i) \dots P(x_1|\mathbf{pa}_1)$ (1.7)

There are two cases:
<u>*Case 1*</u>: for this combination of values:
$P(x_i, x_{i-1}, \dots, x_1) = 0$ (1.8)
Clearly (1.8) implies
$P(x_{i+1}, x_i, \dots, x_1) = 0$

Furthermore, due to (1.8) and the induction hypothesis, there is some k, where $1 \leq k \leq i$ such that $P(x_k|\mathbf{pa}_k) = 0$. So (1.7) holds.

<u>*Case 2*</u>: For this combination of values:
$P(x_i, x_{i-1}, \dots, x_1) \neq 0$
In this case

$$\begin{aligned} P(x_{i+1}, x_i, \dots, x_1) &= P(x_{i+1}|x_i, \dots, x_1)P(x_i, \dots, x_1) \\ &= P(x_{i+1}|\mathbf{pa}_{i+1})P(x_i, \dots, x_1) \\ &= P(x_{i+1}|\mathbf{pa}_{i+1})P(x_i|\mathbf{pa}_i) \dots P(x_1|\mathbf{pa}_1) \end{aligned}$$

The first equality is due to the rule for conditional probability, the second is due to the Markov condition and the third one is due to the induction hypothesis.

Theorem 1.5 Let a DAG $\mathbb{G}$ be given in which each node is a random variable, and let a discrete conditional probability distribution given the values of its parents in $\mathbb{G}$ be specified. Then the product of these conditional distributions yields a joint probability distribution P of the variables and $(\mathbb{G}, P)$ satisfies the Markov condition.
Proof: Order the nodes according to the ancestral ordering. Let $X_1, X_2, \dots, X_n$ be the resultant ordering. Next define:
$P(x_1, x_2, \dots, x_n) = P(x_n|\mathbf{pa}_n)P(x_{n-1}|\mathbf{pa}_{n-1}) \dots P(x_2|\mathbf{pa}_2)P(x_1|\mathbf{pa}_1)$

where $\mathbf{PA}_i$ is the set of parents of X_i in $\mathbb{G}$ and $P(x_i|\mathbf{pa}_i)$ is the specified conditional probability distribution. First, we show that this does indeed yield joint probability distribution. Clearly, $0 \leq P(x_1, x_2, \ldots, x_n) \leq 1$ for all values of the variables. Therefore to show that we have joint distribution Definition 1.8 and Theorem 1.3 imply that we only need to show that the sum of $P(x_1, x_2, \ldots, x_n)$, as the variables are ranging through their all possible values, equal to 1. To that end:

$\sum_{x_1}\sum_{x_2}\ldots\sum_{x_{n-1}}\sum_{x_n} P(x_1, x_2, \ldots, x_n) =$
$\sum_{x_1}\sum_{x_2}\ldots\sum_{x_{n-1}}\sum_{x_n} P(x_n|\mathbf{pa}_n)P(x_{n-1}|\mathbf{pa}_{n-1})\ldots P(x_2|\mathbf{pa}_2)P(x_1|\mathbf{pa}_1) =$
$\sum_{x_1}\left[\sum_{x_2}\left[\ldots\left[\sum_{x_{n-1}}\left[\sum_{x_n} P(x_n|\mathbf{pa}_n)\right] P(x_{n-1}|\mathbf{pa}_{n-1})\right]\ldots\right]P(x_2|\mathbf{pa}_2)\right]P(x_1|\mathbf{pa}_1) =$
$\sum_{x_1}\left[\sum_{x_2}\left[\ldots\left[\sum_{x_{n-1}}[1]\, P(x_{n-1}|\mathbf{pa}_{n-1})\right]\ldots\right]P(x_2|\mathbf{pa}_2)\right]P(x_1|\mathbf{pa}_1) =$
$\sum_{x_1}\left[\sum_{x_2}[\ldots 1 \ldots]P(x_2|\mathbf{pa}_2)\right]P(x_1|\mathbf{pa}_1) =$
$\sum_{x_1}[1]P(x_1|\mathbf{pa}_1) = 1.$

To be done: *show that the specified conditional distributions are the conditional distributions which they notationally represent in the joint distribution*

Finally, we show the Markov condition is satisfied. To do this, we need to show for $1 \leq k \leq n$ that whenever $P(\mathbf{pa}_k) \neq 0$ if $P(\mathbf{nd}_k|\mathbf{pa}_k) \neq 0$ and $P(x_k|\mathbf{pa}_k) \neq 0$ then $P(x_k|\mathbf{nd}_k, \mathbf{pa}_k) = P(x_k|\mathbf{pa}_k)$, where $\mathbf{ND}_k$ is the set of non-descendants of X_k in $\mathbb{G}$. Since $\mathbf{PA}_k \subseteq \mathbf{ND}_k$, we only need to show that $P(x_k|\mathbf{nd}_k) = P(x_k|\mathbf{pa}_k)$. First, for a given k, order the nodes so that all and only descendants of X_k precede X_k in the ordering. Note that this ordering depends on k whereas the ordering in the first part of the proof does not. Clearly then:

$\mathbf{ND}_k = \{X_1, X_2, \ldots, X_{k-1}\}$

Let

$\mathbf{D}_k = \{X_{k+1}, X_{k+2}, \ldots, X_n\}$

In what follows $\sum_{d_k}$ ... means the sum as the variables in $\mathbf{d}_k$ go over all of their possible values. Furthermore, notation such as $\hat{x}_k$means the variable has a particular value; notation such as $\widehat{\mathbf{nd}}_k$ means all variables in the set have particular values; and notation such as $\mathbf{pa}_k$ means some variables in the set may not have particular values. We have that:

$$P(\hat{x}_k|\widehat{\mathbf{nd}}_k) = \frac{P(\hat{x}_k, \widehat{\mathbf{nd}}_k)}{P(\widehat{\mathbf{nd}}_k)}$$
$$= \frac{\sum_{\mathbf{d}_k} P(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k, x_{k+1}, \ldots, x_n)}{\sum_{\mathbf{d}_k \cup \{x_k\}} P(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{k-1}, x_k, \ldots, x_n)}$$
$$= \frac{\sum_{\mathbf{d}_k} P(x_n|\mathbf{pa}_n)\ldots P(x_{k+1}|\mathbf{pa}_{k+1})P(\hat{x}_k|\widehat{\mathbf{pa}}_k)\ldots P(\hat{x}_1|\widehat{\mathbf{pa}}_1)}{\sum_{\mathbf{d}_k \cup \{x_k\}} P(x_n|\mathbf{pa}_n)\ldots P(x_k|\mathbf{pa}_k)P(\hat{x}_{k-1}|\widehat{\mathbf{pa}}_{k-1})\ldots P(\hat{x}_1|\widehat{\mathbf{pa}}_1)}$$
$$= \frac{P(\hat{x}_k|\widehat{\mathbf{pa}}_k)\ldots P(\hat{x}_1|\widehat{\mathbf{pa}}_1)\sum_{\mathbf{d}_k} P(x_n|\mathbf{pa}_n)\ldots P(x_{k+1}|\mathbf{pa}_{k+1})}{P(\hat{x}_{k-1}|\widehat{\mathbf{pa}}_{k-1})\ldots P(\hat{x}_1|\widehat{\mathbf{pa}}_1)\sum_{\mathbf{d}_k \cup \{x_k\}} P(x_n|\mathbf{pa}_n)\ldots P(x_k|\mathbf{pa}_k)}$$
$$= \frac{P(\hat{x}_k|\widehat{\mathbf{pa}}_k)[1]}{[1]} = P(\hat{x}_k|\widehat{\mathbf{pa}}_k)$$

In the second to last step, the sums are each equal to one for the following reason. Each is a sum of a product of conditional probability distributions specified for a DAG. In the case of the numerator, that DAG is the subdigraph of our original digraph $\mathbb{G}$, consisting of the variables in $\mathbf{D}_k$, and in the case of the denominator, it is the subdigraph consisting of the variables in $\mathbf{D}_k \cup \{ X_k \}$. Therefore, the fact that each of those sums equals 1 follows from the first part of the proof. Notice that the theorem requires that the specified conditional distributions be discrete. Often in the case of continuous distributions it still holds. For example it holds for Gaussian distributions. However, it does not hold for all continuous conditional distributions. See [Dawid and Studeny, 1999] for example in which no joint distribution having the specialized distributions as conditionals even exist.

Bayesian Networks

Definition (*Bayesian Network*)
Let P be the joint probability distribution of the random variables in some set V, and $\mathbb{G} = (V, E)$ be a DAG. We call $(\mathbb{G}, P)$ a Bayesian network if $(\mathbb{G}, P)$ satisfies the Markov condition. Owning to *Theorem 1.4*, P is the product of its conditional distributions in $\mathbb{G}$ and this is the way P is always represented in a Bayesian network. Furthermore, owing to *Theorem 1.5*, if we specify a DAG $\mathbb{G}$ and any discrete conditional distributions (and many continuous ones), we obtain a Bayesian network. This is the way Bayesian networks are constructed in practice.

Creating Bayesian Network using Causal Edges

Definition (*causal DAG*) Given a set of random variables V, if for every X, Y in V we draw an edge from X to Y if and only if X is a direct cause of Y relative to V, we call the resulting DAG a ***causal DAG***.

Ascertaining Causal Influences Using Manipulation
Some of what follows is based on a similar discussion to [Cooper, 1999].

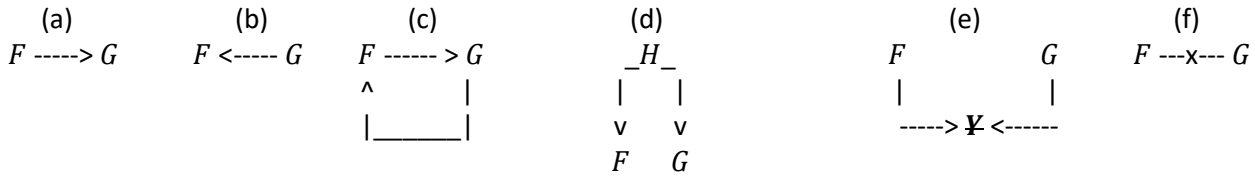
Definition (*Operational method for identifying causal relationships*)
If the action of making some variable X take some value sometimes changes the value taken by variable Y, then we assume X is responsible for sometimes changing Y's value, and we conclude X is a **cause** of Y. More formally, we say we **manipulate** X when we force X to take some value and we say X causes Y if there is some manipulation of X which leads to a change in the probability distribution of Y.

We assume that if manipulating X leads to a change of the probability distribution of Y, then X obtaining a value by any means also leads to a change of the probability distribution of Y. So we assume that causes and their effects are statistically correlated. Note that in general variables can be statistically correlated without one being the cause of the other. A manipulation consists of a randomized controlled experiment (**RCE**) using some specific population of entities in some specific context. The causal relationship discovered is then relative to this population and this context.

Let us discuss how manipulation proceeds. We first identify the population of entities we wish to consider. Our random variables are features of these entities. Next, we ascertain the causal relationship we wish to investigate. Suppose we are trying to determine if variable X is the cause of variable Y. For every entity selected, we manipulate the value of X so that each of its possible values is given to the same number of entities (if X is continuous, we choose the values of X according to uniform distribution). After the value of X is set for a given entity we measure the value of Y for that entity. The more the resultant data shows a dependency between X and Y the more the data supports that X casually influences Y. The manipulation of X can be represented by a variable M that is external to the system being studied. There is one value mi of M for each value xi of X, the probabilities of all values of M are the same, and when M equals mi, X equals xi. That is, the relationship between M and X is deterministic. The data supports that X causally influences Y to the extent the data indicates $P(yi|mi) \neq P(yi|mk)$ for $y \neq k$. Manipulation is actually a special kind of causal relationship that we assume exists primordially and is within our control so that we can define and control other causal relationships.

Example (*Possible causal relationships*)

Let F and G be random variables. The actual values of F and G are unimportant to the current discussion. We could use either continuous or discrete values. If F caused G then, indeed, they would be statistically correlated but this would be the case if G caused F, or if they had some hidden common cause H. If we represent causal influence by a directed edge we have the following 5 possibilities:



(a) Shows the conjecture that F causes G

(b) Shows the conjecture that G causes F

When we do not have domain knowledge (a) and (b) seem equally reasonable.

(c) shows causal loop or feedback.

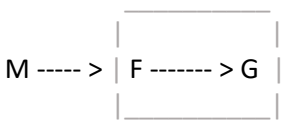
(d) F and G have some hidden common cause H which accounts for their statistical correlation.

(e) we are observing a population in which all individuals have some (possibly hidden) effect of both F and G. We say a node is **instantiated** when we know its value for the entity currently being modeled. So we are saying Y is instantiated to the same value for all entities in the population we are observing. This is depicted here by putting the node ~~**Y**~~ in ~~**bold strikethrough**~~. Ordinarily, an instantiation of a common effect causes a dependency between its causes because each cause explains away the occurrence of the effect, thereby making the other cause less likely. This psychological phenomenon is called **discounting**. So if this were the case discounting would explain the correlation between F and G. This type of dependency is called **selection bias**.

(f) F and G are not related causally at all. The most notable example is when our entities are points in time and our random variables are values of properties at these different points in time. Such random variables are often correlated without having an apparent causal connection.

It may not be obvious why two variables with common cause would be correlated. Consider the present example. Suppose that H is a common cause of F and G and neither F nor G caused the other. Suppose H is a common cause of F and G and neither F nor G caused the other. Then H and F are correlated because H causes F, H and G are correlated because H causes G, which implies that F and G are correlated transitively through H. Here is a more detailed explanation: for this example suppose $h1$ is a value of H that has causal influence on F taking value $f1$ and on G taking value $g1$. Then if F had value $f1$, each of its causes would become more probable because one of them should be responsible. So $P(h1|f1) > P(f1)$. Now since the probability of $h1$ has gone up, the probability of $g1$ would also go up because h1 causes g1. Therefore, $P(g1|f1) > P(f1)$, which means F and G are correlated.

Example (The company $\mathcal{C}$'s *manipulation study*)



$P(m1) = 0.5$
$P(m2) = 0.5$

$P(f1|m1) = 1$
$P(f2|m1) = 0$
$P(f1|m2) = 0$
$P(f2|m2) = 1$

Since $\mathcal{C}$ cannot conclude that F causes G which is remediation of the disease $\mathcal{D}$ from their mere correlation alone they did a test manipulation to test this conjecture. The study was done on N men in a particular age group which exhibited the symptom of the disease $\mathcal{D}$ which F is supposed to cure. $N/2$ of the men were given F and $N/2$ were given placebo. Let us define the variables for the study including the manipulation variable.

Variable	Value	When The Variable Takes This Value
F	$f1$	subject takes given quantity of the substance $\mathcal{F}$
	$f2$	subject takes given quantity of placebo
G	$g1$	subject no longer experiences symptoms of $\mathcal{D}$
	$g2$	subject still experiences symptoms of $\mathcal{D}$
M	$m1$	subject is chosen to take given quantity of the substabce $\mathcal{F}$
	$m2$	subject is chosen to take given quantity of placebo

The figure above shows the conjecture that F causes G and the **RCE** used to test this conjecture. The gray line around the system being modeled indicates that the manipulation comes from outside the system. The edges in that graph represent causal influences. The **RCE** supports the conjecture that F causes G to the extent that the data supports $P(g1|m1) \neq P(g1|m2)$. Specifically, it was found that $P(g1|m1) = 0.67$ and $P(g1|m2) = 0.07$.

Example (*Causal Mediaries*)

F ---- > A ---- > G

A – causal mediary

Let us suppose that there is an agent $\mathcal{A}$ accounted by another random variable A such that F and A are in causal relationship. Let us suppose that $\mathcal{C}$ had enough information to conclude that A and G are in causal relationship as well. These two causal relationships are depicted on the Graph above.
Could $\mathcal{C}$ have assumed that F had a causal effect on G through A and thus avoiding to do the manipulation **RCE**? The answer is No. It is possible that certain minimal level of the agent $\mathcal{A}$ is necessary to trigger the disease $\mathcal{D}$, more than that minimal level of $\mathcal{A}$ has no further effect on $\mathcal{D}$ and the substance $\mathcal{F}$ is not capable of lowering the level of $\mathcal{A}$ beyond that minimal level. That is, it may be that $\mathcal{F}$ has causal effect on $\mathcal{A}$ and $\mathcal{A}$ has causal effect on $\mathcal{D}$ and yet $\mathcal{F}$ has no causal effect on $\mathcal{D}$.

Definition (*faithfulness condition*) If we identify that F causes A and A causes G, and F and G are probabilistically independent we say that the probability distribution of the variables is not **faithful** to the DAG representing their causal relationships. In general, we say $(\mathbb{G}, P)$ satisfies the **faithfulness condition** if $(\mathbb{G}, P)$ satisfies Markov condition and the only conditional independencies are entailed by the Markov condition. So if F and G are independent, the probability distribution does not entail the faithfulness condition in the DAG of the figure above because this independence is not entailed by the Markov condition.

Notice that the variable A was not in the DAG on the figure above and if probability distribution did satisfy the faithfulness condition there would have been an edge from F directly to G instead of taking the directed path through A. It seems that we can usually conceive of intermediate unidentified variables along each edge. Consider the following example:

Example (*causal mediary*): If C is an event of striking a match, and A is an event of the match catching on fire, and no other events are considered then C is a direct cause of A. If however, we added B, the sulfur on the match achieved sufficient heat to combine with oxygen then we could no longer say that C directly caused A but rather C caused B and B caused A. Accordingly, we say that B is a **causal mediary** between C and A if C causes B and B causes A.

Note (*observer-dependent variables*): In that intuitive explanation a variable name is used to stand also for the value of the variable. For example, A is a variable whose value is on-fire or not-on-fire and A is also used to represent that the match is on fire. Clearly, we can add more causal mediaries. For example, we could add the variable D representing whether the match tip is abraded by a rough surface. C, then, would cause D, which in turn would case B, etc. We see that the set of **observable** variables are **observer-dependent**. An individual, given large amount of sensory input, selectively records discernible events and develops cause-effect relationships between them. Therefore, rather than assuming that there is an objective set of causaly related variables , it is more appropriate to assume that in the given context of the application we identify only certain variables and develop a set of causal relationships between them.

Bad Manipulation

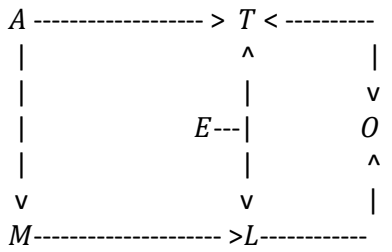
Before discussing causation and the Markov condition, we note some cautionary procedures of which one must be aware when performing an **RCE**. First, we must be careful that we do not inadvertently disturb the system other than the disturbance done by manipulating the variable M itself. That is we must be careful we do not introduce more causal edges in the system being modeled.

Example (*Bad Manipulation*):

Suppose we want to determine the relative effectiveness of home treatment and hospital treatment for low-risk pneumonia patients. Consider those patients of Dr X who are randomized for home treatment but whom should have been normally admitted to hospital. Dr X may give more instructions to such home-bound patients than he would give to the ***typical*** home bound patient. These instructions might influence patient outcomes. If those instructions are not measured, then the **RCE** may give biased estimates of the effect of the treatment location (home or hospital) on patient outcome. Note, we are interested in estimating the effect of treatment location on patient outcomes, everything else being equal. The **RCE** is actually telling us the effect of treatment allocation on patient outcomes, which is not of interest here. The manipulation of treatment location is a bad manipulation because it not only results in manipulation M of treatment location but also has causal effect on physican's other actions such as advice given. This is an example of what is called ***fat-hand manipulation*** in the sense that one wants to manipulte just one variable but one's hand is so fat so it ends up manipulating other variables. Let us show with a DAG how this **RCE** inadvertently disturbs the system being modeled other than the disturbance done by M itself. If we let L represent a treatment location, A represent treatment allocation and M represent the manipulation of treatment location.

Variable	Value	When The Variable Takes This Value
L	$l1$	subject is at home
	$l2$	subject is at hospital
A	$a1$	subject is allocated to be at home
	$a2$	subject is allocated to be at hospital
M	$m1$	subject is chosen to stay home
	$m2$	subject is chosen to stay at the hospital

Other variables of the system:
E – doctor's evaluation of the patient
T – doctor's treatment
O – patient outcome
Since the last 3 variables can have more than two values we did not show them in the table above.



Causation and the Markov condition

Let us have a causal DAG $G(\mathcal{V},\mathcal{A})$. This means that given a set of variables $\mathcal{V}$, if there is an edge between X and Y then X is the direct cause of Y relative to $\mathcal{V}$. Then a manipulation of X would change the probability distribution of Y in a such way that there would be no subset $W \subseteq V - \{X, Y\}$ such that if we instantiate the variables in W a manipulation of X no longer changes the probability distribution of Y.

When constructing a causal DAG containing a set of variables V, we call V our ***set of observed variables***.

Why Causal DAGs Often Satisfy The Markov Condition

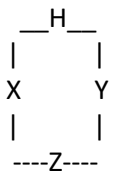
Let us consider the earlier example on company $\mathcal{C}$'s manipulation study.
In this case the set of observed variables $\mathcal{V}$ is $\{F, D, G\}$. We do have a causal edge from F to D as it was shown on the Figure. We also do have causal edge from D to G. We suspect that F influences G only through D so we did not place a causal edge from F to G. If there is another causal path from F to G (i.e. F affected G by some means other than D we would also place an edge from F to G. Assuming the only causal connection between F and G is as indicated on the Figure it appears that F and G are conditionally independent given D. This conditional independence holds because once we knew the value of D we would have the probability distribution of G based on this known value and since F cannot change the known value of D and there is no other connection between F and G it cannot change the probability distribution of G. Manipulation experiments have substantiated this intuition. That is, there have been experiments in which it was established that X causes Y, Y causes Z, X and Z are not probabilistically independent and X and Z are conditionally independent given Y.

Theorem (*informal statement for Markov condition validity in causal DAGs*): In general, when all causal paths from X to Y contain at least one variable in our set of observed variables V, X, and Y do not have common cause, there are no causal paths from Y back to X, and we do not have selection bias then X and Y are independent if we condition on a set of variables including at least one variable in each of the causal paths from X to Y.

Definition (*common cause*): we say that X and Y have **common cause** if there is some variable that has causal paths into both X and Y. If X and Y have a common cause C, there is often a dependency between them through that common cause. However, if we condition on Y's parent in the path from C to Y, we can break this dependency for the same reasons discussed above. So as long as all common

causes are in our set of observed variables V, we can still break the dependency between X and Y (assuming there are no causal paths from Y to X) by conditioning on the set of parents of Y, which means the Markov condition is still satisfied relative to X and Y.

Definition (*hidden variable*): A problem arises when at least one common cause is not in our set of observed variables V. Such common cause is called hidden variable. If two variables had a hidden common cause, then there would often be a dependency between them, which the Markov condition would identify as independency. For example, consider the DAG shown below:



If we only identified the variables X, Y and Z and causal relationships that X and Y each caused Z, we would draw edges from each of X and Y to Z. The Markov condition would entail X and Y are independent. But if X and Y had hidden common cause H, they would not ordinarily be independent. So, for us to assume the Markov condition is satisfied, either no two variables in the set of observed variables V can have a hidden common cause, or, if they do, it must have the same unknown value for every unit in the population under consideration.

Definition (*causally sufficient*): When the observed variables have a hidden common cause which impacts by the same unknown value every variable in the population we say the observed variables are **causally sufficient**.

Another violation of the Markov condition, similar to the failure to include a hidden common cause is when there is a selection bias present.
Definition (*selection bias*)
Let us consider again the earlier example on company $\mathcal{C}$'s manipulation study. If the medication F and apprehension that the test subject is using the medication F both lead to Y (hypertension) and we are observing individuals hospitalized for hypertension we would observe probabilistic dependence between F and G due to **selection bias**. This is shown again on the figure below:

Note that in this situation our set of observed variables is $\{F, G\}$. That is, Y is unobserved. So, if neither F nor G caused each other and they did not have hidden common cause a causal DAG containing only two variables (i.e. with no edges) would still not satisfy the Markov condition with the observed probability distribution because the Markov condition says F and G are independent when indeed they are not for this population.

Finally, we must make sure that if X has causal influence on Y, then Y does not have causal influence on X. In this way we guarantee that the identified causal edges will indeed yield a DAG. Causal feedback loops are discussed in [Richardson and Sprites, 1999].

One final remark, if we mistakenly draw an edge from X to Y in a case where X's causal influence on Y is only through other variables in the model, we have not done anything to thwart the Markov condition being satisfied. For instance, consider again the Figure:

F ----- > D ------ > G

If F's only influence on G is through D, we would not thwart the Markov's condition by drawing an edge from F to G. That is, this does not result in the structure of DAG entailing any conditional dependencies which are not there. Instead the opposite has happened – the DAG fails to entail conditional independency (namely $I(\{F\},\{G\}|\{D\})$) that is there. This is violation of the *faithfulness condition*, not the *Markov condition*. In general, we would not want to do that because it makes the DAG less informative and unnecessarily increases the size of the instance which is important because the problem of doing Bayesian inference is NP-complete.

The Causal Markov Assumption

We've offered a definition of causation based on manipulation and we've argued that, given this definition of causation, a causal DAG often satisfies the Markov condition with the probability distribution of the variables which means we can construct a Bayesian network by creating a causal DAG.

Definition (*causal Markov assumption*) we say we are making causal Markov assumption if we create causal DAG $\mathcal{G} = (V, E)$ and assume that the probability distribution of the variables in V satisfies the Markov condition with $\mathcal{G}$.
As discussed above, if the following three conditions are satisfied the causal Markov assumption is ordinarily warranted:

1) There must be no *hidden common causes*
2) Selection bias must not be present
3) There must be no causal feedback loops

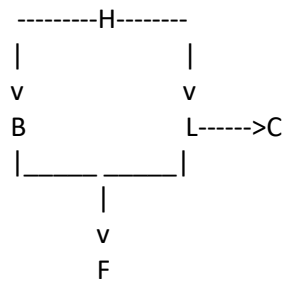
In general, when constructing a Bayesian network using identified causal influences, one must take care that the causal Markov assumptions hold.

Note: We often identify causes using methods other than manipulation. For example, most of us believe smoking causes lung cancer. Yet, we have not manipulated individuals by making them smoke. We believe in this causal influence because smoking and lung cancer are correlated, the smoking precedes the cancer in time (common assumption is that an effect cannot precede the cause) and there are biochemical changes associated with smoking. All of this could possibly be explained by hidden common cause but domain experts rule out this possibility. When we identify causes by any means our belief is that they can be identified by manipulation if we were to perform **RCE** and we make causal Markov assumption as long as we are confident that 1), 2) and 3) are not present.

Example (*cause identification based on elimination of 1), 2), and 3)*):
Suppose we have identified the following causal influences by some means: history of smoking (H) has a causal effect both on bronchitis (B) and on lung cancer (L). Furthermore, each of these variables can cause a fatigue (F). Lunger Cancer (L) can cause positive chest X-Ray (C). Then the DAG on the Figure below represents our identified causal relationships among these variables. If we believe that:

1) These are the only causal influences among the variables
2) There are no hidden common causes
3) Selection bias is not present

it seems reasonable to make the causal Markov assumption.



The Markov Condition Without Causation

Using causal edges is just one way to develop a DAG and a probability distribution that satisfy the Markov condition. In a previous example we showed the joint distribution of V (value), S (shape), and C (color) satisfied the Markov condition with the DAG below but we would not say that the color of an object has a causal influence on its shape. The Markov condition is simply a property of the probabilistic relationship between the variables. Furthermore, if the DAG on the Figure (a) below did capture the causal relationships among some causally sufficient set of variables and there was no selection bias present, the Markov condition would be satisfied not only in (a) but also in (b) and (c). Yes we certainly would not say that the edges in (b) and (c) represent causal influence.

(a) (b) (c)

V <------ C -------> S V ------> C -------> S V <------ C <------- S

Example (*first Example on Markov condition satisfied by causal DAG*)

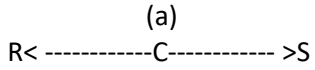
If Alice's husband Ralph was planning a surprise birthday party for Alice with a caterer (C), this may cause him to visit the caterer's store (V). The act of visiting that store would cause him to be seen (S) visiting that store. The causal relationships between the variables are like the ones depicted in Figure (a) below. There is no direct path from C to S because planning the party with the caterer could only cause him to be seen visiting the store *if it caused him to actually visit the store*. If Alice's friend Trixie reported to her that she had seen Ralph visiting the caterer's store today, Alice would conclude that he may be planning a surprise birthday party because she would feel there is a good chance Trixie really did see Ralph visiting the store, and in this case there is a chance he maybe planning a surprise birthday party. So C and S are not independent. If, however, Alice has witnessed this same act of Ralph visiting the caterer's store, she would already suspect Ralph may be planning surprise birthday party. Trixie's testimony would not affect her belief concerning Ralph's visiting the store and therefore would have no affect on her belief concerning his planning a party. So C and S are conditionally independent given V as the Markov condition entails for the DAG on Figure (a). The instantiation of V which renders C and S independent is depicted on Figure (b).

(a) (b)

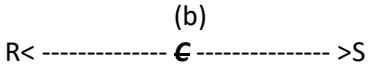
C ------ >V -------> S C ------> **V** -------> S

Example (*second Example on Markov condition satisfied by causal DAG*)
A cold (C) can cause both sneezing (S) and runny nose (R).

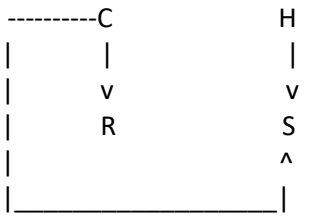
(a)



(b)



(c)



(d)