CONDITIONAL INDEPENDENCE AND GRAPHOIDS

Definition 1.1.2 (Conditional Independence) *Let $V = \{V_1, V_2, \ldots\}$ be a finite set of variables. Let $P(\cdot)$ be a joint probability function over the variables in V, and let X, Y, Z stand for any three subsets of variables in V. The sets X and Y are said to be* **conditionally independent** *given Z if*

$$P(x|y,z) = P(x|z) \quad \textit{whenever} \quad P(y,z) > 0. \tag{1.26}$$

In words, learning the value of Y does not provide additional information about X, once we know Z.

Interpretation:

For any configuration x of the variables in the set X and for any configurations y and z of the variables in Y and Z satisfying $P(Y = y,\ Z = z) > 0$, we have

$$P(X = x|Y = y,\ Z = z) = P(X = x|Z = z). \tag{1.27}$$

CONDITIONAL INDEPENDENCE
(Notation)

$$(X \perp\!\!\!\perp Y | Z)_P \quad \text{iff} \quad P(x|y,z) = P(x|z) \qquad \textbf{(1.28)}$$

for all values x, y, z such that $P(y,z) > 0$.

Special case: Marginal independence

$$(X \perp\!\!\!\perp Y | \emptyset) \quad \text{iff} \quad P(x|y) = P(x) \quad \text{whenever} \quad P(y) > 0 \qquad \textbf{(1.29)}$$

THE GRAPHOID AXIOMS

1. *Symmetry:* $(X \perp\!\!\!\perp Y | Z) \Longrightarrow (Y \perp\!\!\!\perp X | Z)$.
2. *Decomposition:* $(X \perp\!\!\!\perp YW | Z) \Longrightarrow (X \perp\!\!\!\perp Y | Z)$.
3. *Weak union:* $(X \perp\!\!\!\perp YW | Z) \Longrightarrow (X \perp\!\!\!\perp Y | ZW)$.
4. *Contraction:* $(X \perp\!\!\!\perp Y | Z) \,\&\, (X \perp\!\!\!\perp W | ZY)$
   $\Longrightarrow (X \perp\!\!\!\perp YW | Z)$.
5. *Intersection:* $(X \perp\!\!\!\perp W | ZY) \,\&\, (X \perp\!\!\!\perp Y | ZW)$
   $\Longrightarrow (X \perp\!\!\!\perp YW | Z)$.

   (Intersection is valid in strictly positive probability distributions.)

Intuitive interpretation: The irrelevant is irrelevant to relevance

Homework 1, Question 1

Prove that the graphoid properties are satisfied if we interpret $(X \perp\!\!\!\perp Y | Z)$ to mean “all paths from a subset X of nodes to a subset Y of nodes are intercepted by a subset Z of nodes.”

GRAPHICAL NOTATION AND TERMINOLOGY

Graph:

1. A set V of *vertices* (or *nodes*)
2. A set E of *edges* (or *links*)

Vertices: adjacent, connected

Edges: directed, undirected, bidirected

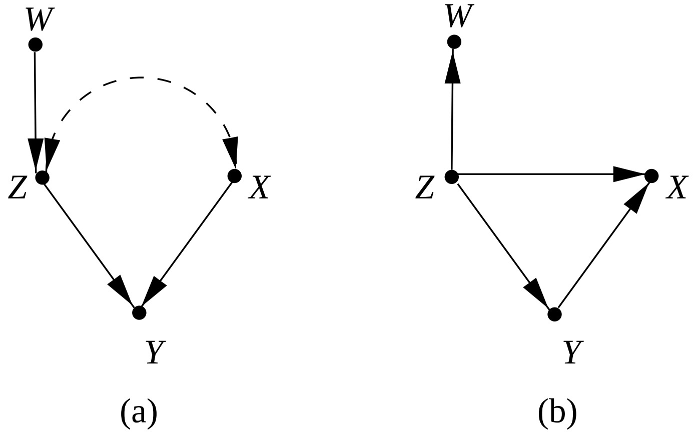

Figure 1.1: (a) A graph containing both directed and bidirected edges. (b) A directed acyclic graph (DAG) with the same skeleton as (a).

Graph concepts: Skeleton, path (unbroken, nonintersecting route traced along the edges in a graph, along or against the arrows), directed path. Cyclic and acyclic graphs, directed acyclic graph (DAG).

Kinship relations: parents, children, descendants, ancestors, spouses (along the full arrows)

Root nodes: (no parents)
Sink nodes (no children)

Every DAG has at least one root and at least one sink.

Tree: A connected DAG in which every node has at most one parent.

Chain: a tree in which every node has at most one child.

Complete graph: every pair of nodes is connected.

BAYESIAN NETWORKS

Graphs in probabilistic reasoning

1. to provide convenient means of expressing substantive assumptions;

2. to facilitate economical representation of joint probability functions; and

3. to facilitate efficient inferences from observations.

BAYESIAN NETWORKS

Directed acyclic graphs emphasizing 3 “Bayesian” aspects

1. The subjective nature of the input information;

2. the reliance on Bayes’s conditioning as the basis for updating information;

3. the distinction between causal and evidential modes of reasoning

DECOMPOSITION BY BAYESIAN NETWORKS

Given a distribution P, on n discrete variables, $X_1, X_2, \ldots, X_n$. Decompose P by the chain rule:

$$P(x_1, \ldots, x_n) = \prod_j P(x_j | x_1, \ldots, x_{j-1}). \qquad \textbf{(1.30)}$$

Suppose X_j is independent of all other predecessors, once we know the value of a select group of predecessors called PA_j. Simplification:

$$P(x_j | x_1, \ldots, x_{j-1}) = P(x_j | pa_j) \qquad \textbf{(1.31)}$$

PA_j : *Markovian parents* of X_j, relative to a given ordering.

Definition 1.2.1 (Markovian Parents)
Let $V = \{X_1, \ldots, X_n\}$ be an ordered set of variables, and let $P(v)$ be the joint probability distribution on these variables. A set of variables PA_j is said to be **Markovian parents** *of X_j if PA_j is a minimal set of predecessors of X_j that renders X_j independent of all its other predecessors. In other words, PA_j is any subset of $\{X_1, \ldots, X_{j-1}\}$ satisfying*

$$P(x_j|pa_j) = P(x_j|x_1, \ldots, x_{j-1}) \qquad \textbf{(1.32)}$$

and such that no proper subset of PA_j satisfies **(1.32).**

Interpretation:
Knowing the values of other preceding variables is redundant once we know the values pa_j of the parent set PA_j.

CONSTRUCTING A BAYESIAN NETWORK

Given: P, and an ordering of the variables.

At the jth stage, select any minimal set of X_j's predecessors that screens off X_j from its other predecessors.

Call this set PA_j, and draw an arrow from each member in PA_j to X_j.

The result is a directed acyclic graph, called a **Bayesian network**, in which an arrow from X_i to X_j assigns X_i as a Markovian parent of X_j, consistent with Definition 1.2.1

The resulting network is unique given the ordering of the variables, whenever the distribution $P(v)$ is strictly positive.

A TYPICAL BAYESIAN NETWORK

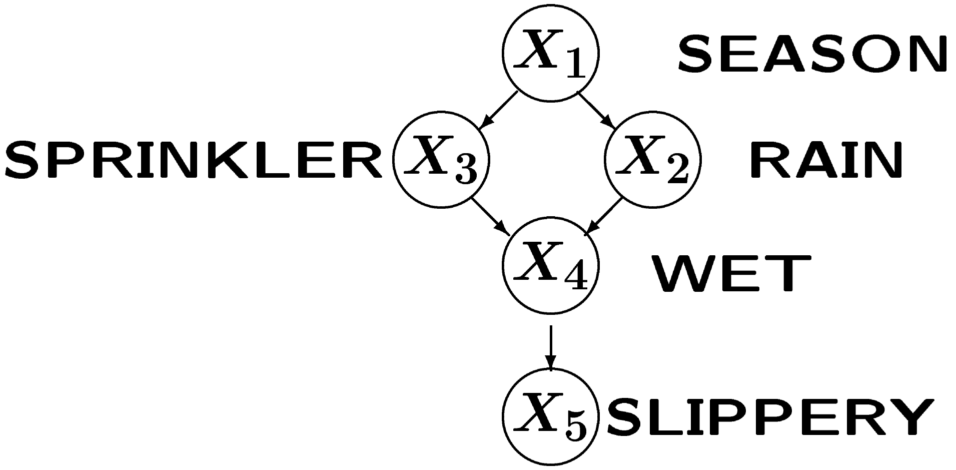

Figure 1.2

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid pa_i) \qquad \textbf{(1.33)}$$

$$\begin{aligned} P(x_1, x_2, x_3, x_4, x_5) \quad &= \quad P(x_1)P(x_2|x_1)P(x_3|x_1) \\ &\quad\;\; P(x_4|x_2, x_3)P(x_5|x_4) \end{aligned} \qquad \textbf{(1.34)}$$

MARKOV COMPATIBILITY

Definition 1.2.2 (Markov Compatibility) *If a probability function P admits the factorization of* (1.33) *relative to DAG G, we say that G* **represents** *P, that G and P are* **compatible**, *or that P is* **Markov relative** *to G.*

Compatibility implies that G can "explain" the generation of the data represented by P.

THE d-SEPARATION CRITERION

Definition 1.2.3 (d-Separation)
A path p is said to be **d-separated** (*or* **blocked**) *by a set of nodes Z if and only if*

1. *p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in Z, or*

2. *p contains an inverted fork (or* **collider***) $i \rightarrow m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.*

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

d-SEPARATION (EXAMPLE)

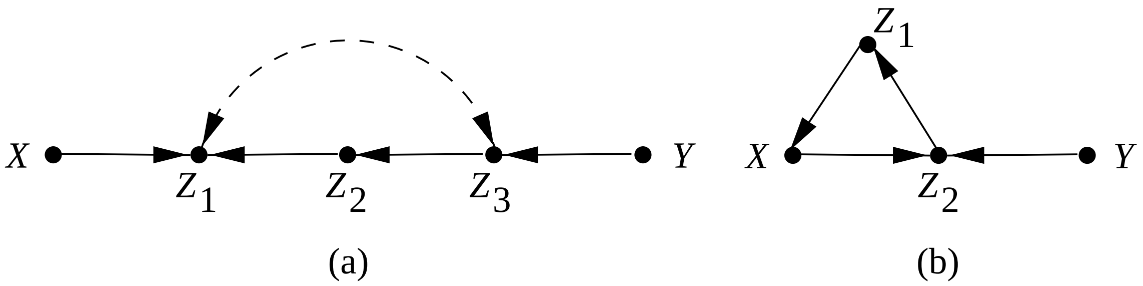

Figure 1.3: Graphs illustrating d-separation. In (a), X and Y are d-separated given Z_2 and d-connected given Z_1. In (b), X and Y cannot be d-separated by any set of nodes.

Theorem 1.2.4
(Probabilistic Implications of d-Separation)
If sets X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every distribution compatible with G. Conversely, if X and Y are **not** *d-separated by Z in a DAG G, then X and Y are dependent conditional on Z in at least one distribution compatible with G.*

Theorem 1.2.5
For any three disjoint subsets of nodes (X, Y, Z) in a DAG G and for all probability functions P, we have:

(i) $(X \perp\!\!\!\perp Y|Z)_G \Longrightarrow (X \perp\!\!\!\perp Y|Z)_P$ *whenever G and P are compatible, and*

(ii) *if $(X \perp\!\!\!\perp Y|Z)_P$ holds in all distributions compatible with G, it follows that $(X \perp\!\!\!\perp Y|Z)_G$.*

Theorem 1.2.6
(Ordered Markov Condition)
A necessary and sufficient condition for a probability distribution P to be Markov relative a DAG G is that, conditional on its parents in G, each variable be independent of all its predecessors in some ordering of the variables that agrees with the arrows of G.

Theorem 1.27
(Parental Markov Condition)
A necessary and sufficient condition for a probability distribution P to be Markov relative a DAG G is that every variable be independent of all its nondescendants (in G), conditional on its parents.

Theorem 1.28
(Observational Equivalence)
Two DAGs are observationally equivalent if and only if they have the same skeletons and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow (Verma and Pearl 1990).

CAUSAL BAYESIAN NETWORKS

Advantages of using causal relations

1. Reliable judgments in network construction (e.g., try $(X_5, X_1, X_3, X_2, X_4)$).

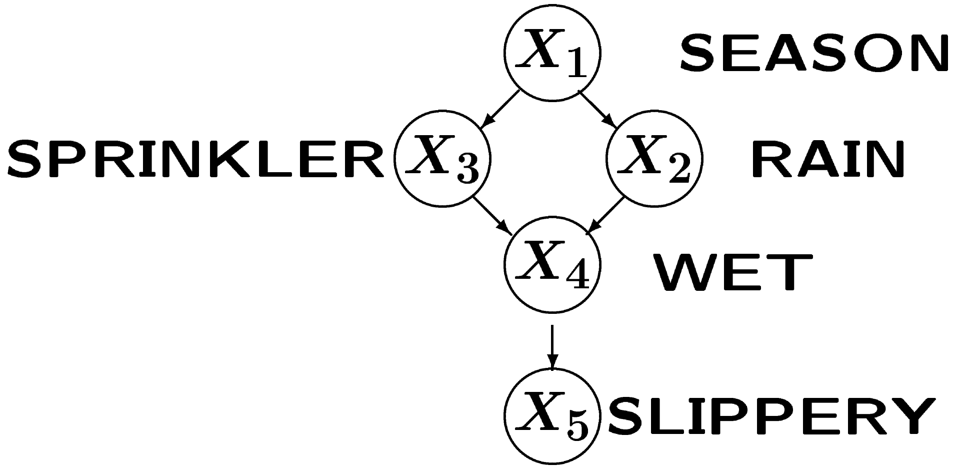

Figure 1.2

Conditional independence judgments are accessible only when they are anchored onto causal relationships.

2. Representation of action and changes (remodeling) (e.g., a disabled sprinkler).
3. Ease of reconfiguration: Modularity, stability and autonomy, (deliberative vs. reactive agents).

CAUSAL NETWORKS: ORACLES FOR INTERVENTIONS

Probabilities do not predict effects of interventions.

connection between modularity and interventions: alteration + simulation

Example:

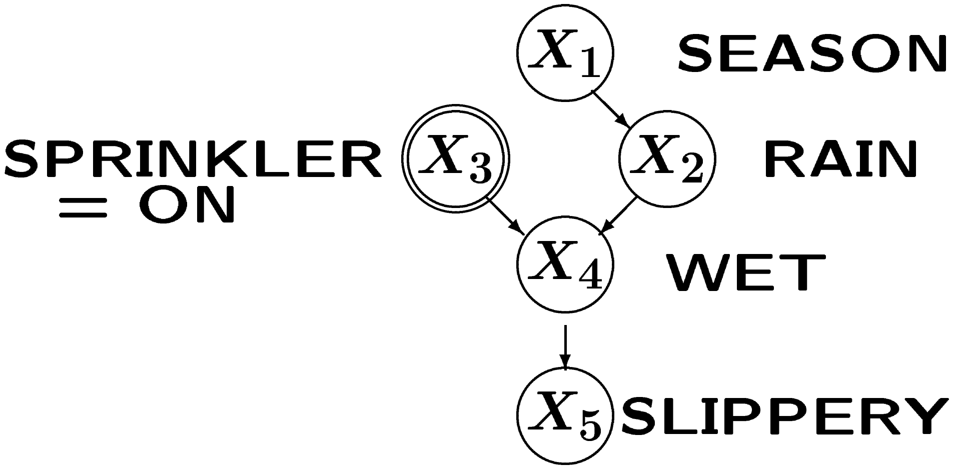

Figure 1.4

$$P_{X_3=\text{On}}(x_1, x_2, x_4, x_5) = P(x_1)P(x_2|x_1) P(X_4|X_2, X_3 = \text{On})P(x_5|x_4) \quad \textbf{(1.36)}$$

ASSUMPTIONS UNDERLYING CAUSAL NETWORKS

Definition 1.3.1 (Causal Bayesian Network) Let $P(v)$ be a probability distribution on a set V of variables, and let $P_x(v)$ denote the distribution resulting from the intervention $do(X = x)$ that sets a subset X of variables to constants x. Denote by $\boldsymbol{P_*}$ the set of all interventional distributions $P_x(v)$, $X \subseteq V$, including $P(v)$, which represents no intervention (i.e., $X = \emptyset$). A DAG G is said to be a **causal Bayesian network** compatible with $\boldsymbol{P_*}$ if and only if the following three conditions hold for every $P_x \in \boldsymbol{P_*}$:

(i) $P_x(v)$ is Markov relative to G;

(ii) $P_x(v_i) = 1$ for all $V_i \in X$ whenever v_i is consistent with $X = x$;

(iii) $P_x(v_i|pa_i) = P(v_i|pa_i)$ for all $V_i \notin X$ whenever pa_i is consistent with $X = x$.

COMPUTING INTERVENTIONAL DISTRIBUTIONS

The distribution $P_x(v)$ resulting from the intervention $do(X = x)$ is given as a **truncated-factorization**

$$P_x(v) = \prod_{\{i|V_i \notin X\}} P(v_i|pa_i) \text{ for all } v \text{ consistent with } x, \qquad \textbf{(1.37)}$$

Properties:

Property 1

For all i,

$$P(v_i|pa_i) = P_{pa_i}(v_i). \qquad \textbf{(1.38)}$$

Property 2

For all i and for every subset S of variables disjoint of $\{V_i, PA_i\}$, we have

$$P_{pa_i,s}(v_i) = P_{pa_i}(v_i). \qquad \textbf{(1.39)}$$

Causal Relationships and Their Stability

Ontological vs. epistemic claims.

S_1: “Turning the sprinkler on would not affect the rain,”

S_2: “Belief in the state of the rain is independent of knowledge of the state of the sprinkler.”

S_2 changes from false to true when we learn what season it is (X_1).

Given X_1, S_2 changes from true to false once we observe that the pavement is wet ($X_4 = \textit{true}$).

S_1 remains true regardless of X_1 or X_4.

Remark: the latter requires counterfactual analysis

INVARIANCE TO PROCESSES

S_1 remains invariant to changes in *all* mechanisms shown in this causal graph.

Causal claims remain invariant to changes in the mechanism that governs the causal variables (X_3 in our example).

Causal claims are sensitive to only those mechanisms that mediate between the cause and the effect.

Probabilistic claims also depends on the context in which those mechanisms are embedded.

FUNCTIONAL CAUSAL MODELS

Compare two specifications:

$$x_i = f_i(pa_i, u_i), \qquad i = 1, \ldots, n, \qquad \textbf{(1.40)}$$

$$P(v_i|pa_i) = P_{pa_i}(v_i). \qquad \textbf{(1.38)}$$

Laplacian vs. stochastic models

1. Many to one
2. In tune with human intuition.
3. Permits counterfactual analysis

e.g., “the probability that event B occurred *because* of event A,” or, “the probability that event B would have been *different* if it were not for event A”

STRUCTURAL EQUATIONS (EXAMPLES)

$$x_i = f_i(pa_i, u_i), \qquad i = 1, \ldots, n, \qquad \textbf{(1.40)}$$

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \qquad i = 1, \ldots, n, \qquad \textbf{(1.41)}$$

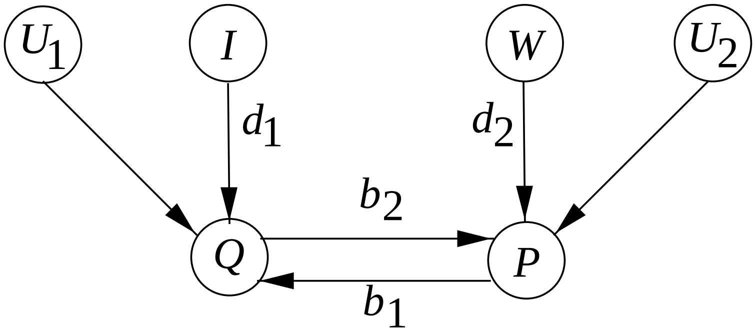

Figure 1.5: Causal diagram illustrating the relationship between price (P), demand (Q), income (Z), and wages (W).

$$\begin{aligned} q &= b_1 p + d_1 i + u_1, && \textbf{(1.42)} \\ p &= b_2 q + d_2 w + u_2, && \textbf{(1.43)} \end{aligned}$$

STRUCTURAL EQUATIONS (EXAMPLES)

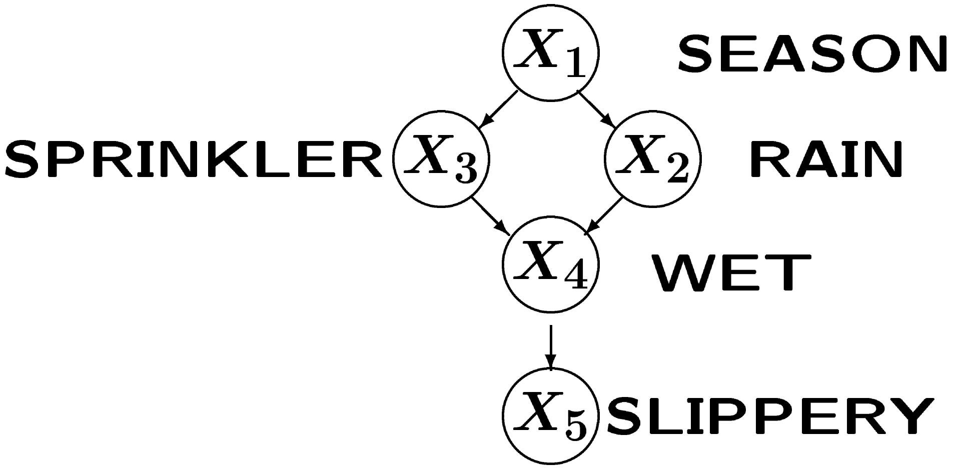

Figure 1.2

$$\begin{aligned} x_1 &= u_1, \\ x_2 &= f_2(x_1, u_2), \\ x_3 &= f_3(x_1, u_3), \\ x_4 &= f_4(x_3, x_2, u_4), \\ x_5 &= f_5(x_4, u_5). \end{aligned} \qquad \textbf{(1.44)}$$

$$\begin{aligned} x_2 &= [(X_1 = \text{winter}) \vee (X_1 = \text{fall}) \vee u_2] \wedge \neg u_2', \\ x_3 &= [(X_1 = \text{summer}) \vee (X_1 = \text{spring}) \vee u_3] \wedge \neg u_3', \\ x_4 &= (x_2 \vee x_3 \vee u_4) \wedge \neg u_4', \\ x_5 &= (x_4 \vee u_5) \wedge \neg u_5', \end{aligned} \qquad \textbf{(1.45)}$$

HIERARCHY OF QUERIES

predictions (e.g., would the pavement be slippery if we *find* the sprinkler off?);

interventions (e.g., would the pavement be slippery if we *make sure* that the sprinkler is off?); and

counterfactuals (e.g., would the pavement be slippery *had* the sprinkler been off, given that the pavement is in fact not slippery and the sprinkler is on?).

PROBABILISTIC PREDICTIONS IN CAUSAL MODELS

Causal model:

$$x_i = f_i(pa_i, u_i), \qquad i = 1, \ldots, n, \qquad \textbf{(1.40)}$$

Causal diagram: draw an arrow from each member of PA_i toward X_i

Semi-Markovian model: acyclic diagram. $P(x_1, \ldots, x_n)$ is determined by $P(u)$

Markovian model: acyclicity and error independence

THE CAUSAL MARKOV CONDITION

Theorem 1.4.1 (Causal Markov Condition) Every Markovian causal model M induces a distribution $P(x_1, \ldots, x_n)$ that satisfies the parental Markov condition relative the causal diagram G associated with M; that is, each variable X_i is independent on all its non-descendants, given its parents PA_i in G (Pearl and Verma 1991)

Proof:
The distribution $P(x_1, \ldots, x_n, u_1, \ldots, u_n)$ is certainly Markov relative the augmented DAG $G(X, U)$, in which the U variables are represented explicitly. The required Markov condition of the marginal distribution $P(x_1, \ldots, x_n)$ follows by d-separation in $G(X, U)$.

IMPLICATIONS OF THE CAUSAL MARKOV CONDITION

The parental Markov condition follows from two causal assumptions:

(1) V contains every variable that is a cause of two or more other variables, and

(2) Reichenbach's common-cause assumption, ("no correlation without causation,") if any two variables are dependent, then one is a cause of the other *or* there is a third variable causing both.

All the probabilistic properties of Bayesian networks follow from these two assumptions, regardless of the choice of functions $\{f_i\}$ and regardless of the error distributions $P(u_i)$.

PROBABILISTIC PREDICTIONS USING MARKOVIAN CAUSAL MODELS

1. Stability: invariance of conditional independencies to f_i and $P(u_i)$

2. Economical specification: e.g., noisy-OR

3. Assumptions simplified: from conditional independence to common causes

4. re-estimation localized.

INTERVENTIONS AND CAUSAL EFFECTS IN FUNCTIONAL MODELS

Representation: replace equations instead of conditional probability factors.

Truncated factorization still valid in Markovian models.

Advantages:

1. Extensions to feedback systems and non-Markovian models.
2. Modification of parameters: meaningful. (functions generate P, conditional probabilities are derived from P.)
3. discrete state probabilities, linear programming
4. context-specific actions and policies. (action sensitive evidence)

COUNTERFACTUALS IN FUNCTIONAL MODELS

$$P(y|x) = 1/2 \qquad \text{for all } x \text{ and } y. \qquad \mathbf{(1.46)}$$

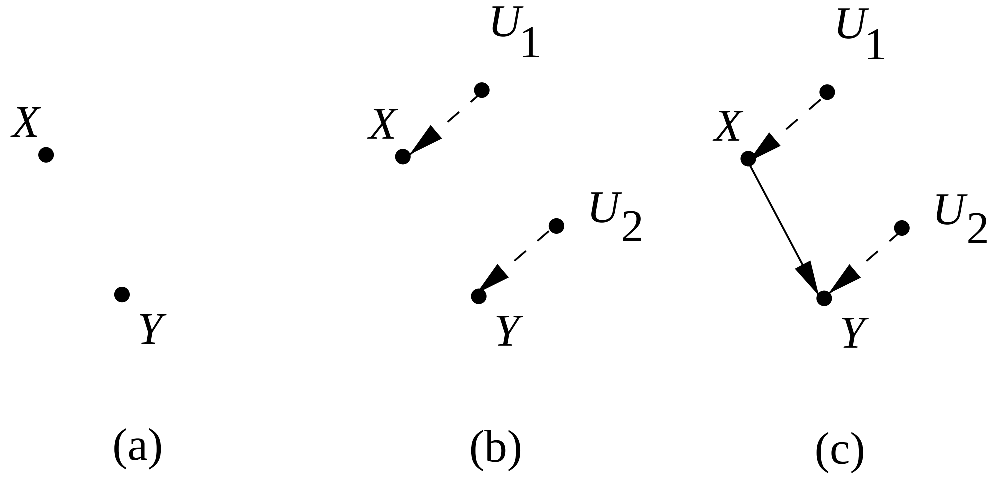

Figure 1.6

Model 1 (Figure 1.6(b))

$$\begin{aligned} x &= u_1, \\ y &= u_2, \end{aligned}$$

Model 2 (Figure 1.6(c))

$$\begin{aligned} x &= u_1, \\ y &= xu_2 + (1-x)(1-u_2), \qquad \mathbf{(1.48)} \end{aligned}$$

CONTINGENCY TABLES FOR MODELS 1 AND 2

Model 1	$u_2 = 0$		$u_2 = 1$		
	$x = 1$	$x = 0$	$x = 1$	$x = 0$	x
$y = 1$ (death)	0	0	0.25	0.25	0
$y = 0$ (recovery)	0.25	0.25	0	0	0

Model 2	$u_2 = 0$		$u_2 = 1$		
	$x = 1$	$x = 0$	$x = 1$	$x = 0$	x
$y = 1$ (death)	0	0.25	0.25	0	0
$y = 0$ (recovery)	0.25	0	0	0.25	0

Figure 1.7

THREE STEPS FOR COMPUTING COUNTERFACTUALS

Given evidence e, compute the probability of $Y = y$ under the hypothetical condition $X = x$ (X and Y subsets of variables)

Step-1 (*abduction*): Update the probability $P(u)$ to obtain $P(u|e)$.

Step-2 (*action*): Replace the equations corresponding to variables in set X by the equations $X = x$.

Step-3 (*prediction*): Use the modified model to compute the probability of $Y = y$.

Step 1 explains the past (U) in light of the current evidence e;

Step 2 bends the course of history (minimally) to comply with the hypothetical condition $X = x$;

Step 3 predicts the future (Y) based on our new understanding of the past and newly established condition, $X = x$.

THREE STEPS FOR COMPUTING COUNTERFACTUALS (example, Model 2)

$$\begin{aligned} x &= u_1, \\ y &= xu_2 + (1-x)(1-u_2), \qquad \textbf{(1.48)} \end{aligned}$$

How do we conclude that a deceased treated subject ($y = 1$, $x = 1$) would have recovered if not treated?

First, we apply the evidence, $e : \{y = 1,\ x = 1\}$, and note that e is compatible with only one realization of U_1 and U_2—namely, $\{u_1 = 1,\ u_2 = 1\}$.

Second, we simulate the hypothetical condition "had he or she not been treated," by substituting $x = 0$ into (1.48) while ignoring the first equation $x = u_1$.

Finally, we solved (1.48) for y (assuming $x = 0$ and $u_2 = 1$) and obtain $y = 0$, from which we conclude that the probability of recovery ($y = 0$) is unity under the hypothetical condition considered.