Probabilistic Graphical Models

Lecture 2: Bayesian Networks

Matthew Walter

TTI-Chicago

April 9, 2020

Some slide content courtesy of David Sontag

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n$

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n$
- How many parameters are required to represent the distribution as a table of probabilities?

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n$
- How many parameters are required to represent the distribution as a table of probabilities? $2^n - 1$ (e.g., $2^{4600} - 1$ for QMR-DT)

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n$
- How many parameters are required to represent the distribution as a table of probabilities? $2^n - 1$ (e.g., $2^{4600} - 1$ for QMR-DT)
- Learning the joint distribution would require a *huge* amount of data

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n$
- How many parameters are required to represent the distribution as a table of probabilities? $2^n - 1$ (e.g., $2^{4600} - 1$ for QMR-DT)
- Learning the joint distribution would require a *huge* amount of data
- Inference of conditional probabilities

$$P(X_i \mid X_j = x) = \frac{P(X_i, X_j = x)}{P(X_j = x)} = \frac{\sum_{X_k \forall k \neq i,j} P(X_1, \ldots, X_n)}{\sum_{X_k \forall k \neq j} P(X_1, \ldots, X_n)}$$

would require summing over exponentially many values

(e.g., 2^{4596} values to determine $P(\text{flu} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$ under QMR-DT)

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n$
- If $P \models \{(X_1 \ldots X_{i-1} \perp X_{i+1} \ldots X_n \mid X_i) \; \forall i\}$,

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n$
- If $P \models \{(X_1 \ldots X_{i-1} \perp X_{i+1} \ldots X_n \mid X_i)\ \forall i\}$,

$$P(x_1, \ldots, x_n) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_2) \ldots P(x_n \mid x_{n-1})$$

then we only need $2n - 1 \ll 2^n - 1$ parameters!

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \dots, X_n$
- If $P \models \{(X_1 \dots X_{i-1} \perp X_{i+1} \dots X_n \,|\, X_i)\ \forall i\}$,

$$P(x_1, \dots, x_n) = P(x_1)P(x_2 \,|\, x_1)P(x_3 \,|\, x_2) \dots P(x_n \,|\, x_{n-1})$$

  then we only need $2n - 1 \ll 2^n - 1$ parameters!
- We need to only learn individual distributions that are much smaller and require far less data
- Inference of conditional probabilities

$$P(X_i \,|\, X_j = x)$$

  would require far fewer summations

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n$
- If $P \models \{(X_1 \ldots X_{i-1} \perp X_{i+1} \ldots X_n \,|\, X_i) \; \forall i\}$,

$$P(x_1, \ldots, x_n) = P(x_1)P(x_2 \,|\, x_1)P(x_3 \,|\, x_2) \ldots P(x_n \,|\, x_{n-1})$$

  then we only need $2n - 1 \ll 2^n - 1$ parameters!
- We need to only learn individual distributions that are much smaller and require far less data
- Inference of conditional probabilities

$$P(X_i \,|\, X_j = x)$$

  would require far fewer summations
- But, there are many distributions that can't be modeled ($n-$dimensional manifold v.s. $2n - 1$ dimensional subspace in $\mathbb{R}^{2^n}$)

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n, Y$
- Naive Bayes: Suppose $P \models \{(X_i \perp \boldsymbol{X}_{-i} \,|\, Y) \; \forall i\}$,

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n, Y$
- Naive Bayes: Suppose $P \models \{(X_i \perp \boldsymbol{X}_{-i} \mid Y) \; \forall i\}$,

$$P(x_1, \ldots, x_n, y) = P(y)P(x_1 \mid y)P(x_2 \mid y) \ldots P(x_n \mid y)$$

then we only need $2n + 1 \ll 2^{n+1} - 1$ parameters!

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n, Y$
- Naive Bayes: Suppose $P \vDash \{(X_i \perp \boldsymbol{X}_{-i} \,|\, Y) \;\forall i\}$,

$$P(x_1, \ldots, x_n, y) = P(y)P(x_1 \,|\, y)P(x_2 \,|\, y) \ldots P(x_n \,|\, y)$$

  then we only need $2n + 1 \ll 2^{n+1} - 1$ parameters!
- We need to only learn individual distributions that are much smaller and require far less data

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \dots, X_n, Y$
- Naive Bayes: Suppose $P \models \{(X_i \perp \boldsymbol{X}_{-i} \,|\, Y) \; \forall i\}$,

$$P(x_1, \dots, x_n, y) = P(y)P(x_1 \,|\, y)P(x_2 \,|\, y) \dots P(x_n \,|\, y)$$

  then we only need $2n + 1 \ll 2^{n+1} - 1$ parameters!
- We need to only learn individual distributions that are much smaller and require far less data
- Inference of conditional probabilities

$$P(X_i \,|\, Y = y)$$

  would require far fewer summations (zero, in fact)

Importance of Independencies (Revisited)

- Consider binary multivariate random variables $X_1, X_2, \ldots, X_n, Y$
- Naive Bayes: Suppose $P \models \{(X_i \perp \boldsymbol{X}_{-i} \mid Y) \; \forall i\}$,

$$P(x_1, \ldots, x_n, y) = P(y)P(x_1 \mid y)P(x_2 \mid y) \ldots P(x_n \mid y)$$

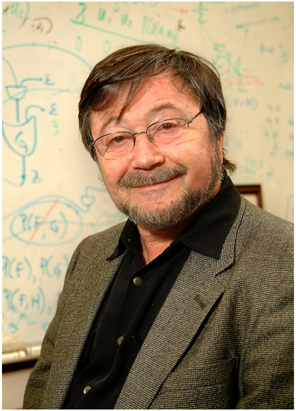
  then we only need $2n + 1 \ll 2^{n+1} - 1$ parameters!
- We need to only learn individual distributions that are much smaller and require far less data
- Inference of conditional probabilities

$$P(X_i \mid Y = y)$$

  would require far fewer summations (zero, in fact)

$\Rightarrow$ Tractable learning and inference requires exploiting independencies

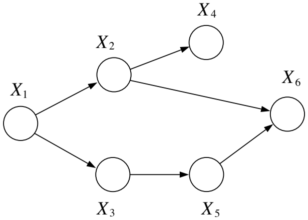
Judea Pearl (1936–)



- First proposed Bayesian networks to encode independence relations c. 1988
- Winner of the 2011 ACM Turing Award for invention of Bayesian networks and algorithms for inference in these models
- Professor at UCLA

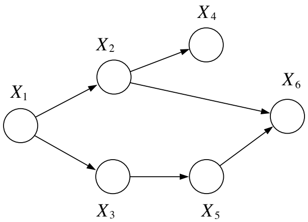
"[Bayesian networks] not only revolutionized the field of artificial intelligence but also became an important tool for many other branches of engineering and the natural sciences." — Turing Award

Bayesian Network Structure



- $G = (V, E)$ is a directed acyclic graph (DAG) s.t.
  - One node $i \in V$ for each random variable X_i
  - $\text{Pa}_{X_i}^G$ denotes the parents of X_i
  - $\text{NonDescendants}_{X_i}$ are variables that are not descendents of X_i

Bayesian Network Structure



- $G = (V, E)$ is a directed acyclic graph (DAG) s.t.
  - One node $i \in V$ for each random variable X_i
  - $\text{Pa}_{X_i}^G$ denotes the parents of X_i
  - $\text{NonDescendants}_{X_i}$ are variables that are not descendents of X_i
- G encodes the following *local* independencies

$$I_l(G) = (X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}^G) \quad \forall X_i$$

i.e., X_i is conditionally independent of $\text{NonDescendants}_{X_i}$ given $\text{Pa}_{X_i}^G$

Bayesian Networks

- A distribution P **factorizes** according to G iff

$$P(X_i, \ldots, X_n) = \prod_{i \in V} P(X_i \,|\, \mathsf{Pa}_{X_i}^G)$$

Bayesian Networks

- A distribution P **factorizes** according to G iff

$$P(X_i, \ldots, X_n) = \prod_{i \in V} P(X_i \,|\, \mathrm{Pa}_{X_i}^{G})$$

- A **Bayesian Network** is a pair $B = (P, G)$ for which
  1. P factorizes over G
  2. P is specified as a set of conditional probability distributions (CPD) $P(X_i \,|\, \mathrm{Pa}_{X_i}^{G})$, one per node specifying probability conditioned on parents

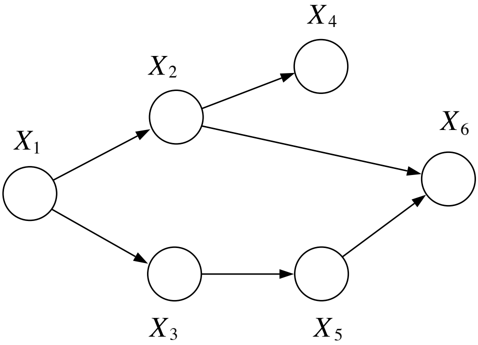
Bayesian Networks

- A distribution P **factorizes** according to G iff

$$P(X_i, \ldots, X_n) = \prod_{i \in V} P(X_i \mid \mathrm{Pa}^G_{X_i})$$

- A **Bayesian Network** is a pair $B = (P, G)$ for which
  1. P factorizes over G
  2. P is specified as a set of conditional probability distributions (CPD) $P(X_i \mid \mathrm{Pa}^G_{X_i})$, one per node specifying probability conditioned on parents
- A graph G is both:
  - a compact representation of the conditional independencies that hold under the corresponding distribution P
  - a data structure that provides a skeleton for compactly representing the joint distribution in a factorized way

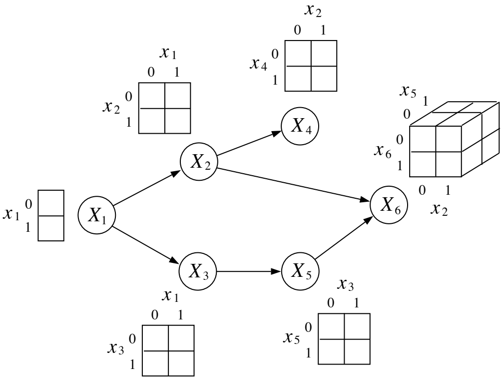
Parametrization and Representation



Representational (storage, learning, & computation) complexity:

- **Joint distribution**: Exponential in the number of variables
- **Bayesian Network**: Exponential in number of parents of each node, linear in the number of nodes

Parametrization and Representation



Representational (storage, learning, & computation) complexity:

- **Joint distribution**: Exponential in the number of variables
- **Bayesian Network**: Exponential in number of parents of each node, linear in the number of nodes

Example: Naive Bayes

- Classify e-mails as spam ($Y = 1$) or not spam ($Y = 0$)
  - Let $1 : n$ index words in a dictionary
  - $X_i = 1$ if word i appears in an e-mail
  - E-mails are drawn according to distribution $P(Y, X_1, \ldots X_n)$

Example: Naive Bayes

- Classify e-mails as spam $(Y = 1)$ or not spam $(Y = 0)$
  - Let $1 : n$ index words in a dictionary
  - $X_i = 1$ if word i appears in an e-mail
  - E-mails are drawn according to distribution $P(Y, X_1, \dots X_n)$
- Suppose words are conditionally independent given Y

$$P(y, x_1, \dots, x_n) = P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

Example: Naive Bayes

- Classify e-mails as spam $(Y = 1)$ or not spam $(Y = 0)$
  - Let $1 : n$ index words in a dictionary
  - $X_i = 1$ if word i appears in an e-mail
  - E-mails are drawn according to distribution $P(Y, X_1, \ldots X_n)$
- Suppose words are conditionally independent given Y

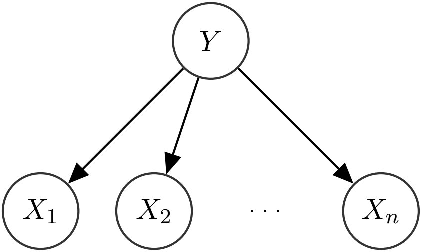
$$P(y, x_1, \ldots, x_n) = P(y) \prod_{i=1}^{n} P(x_i \,|\, y)$$

- Infer (predict) whether an e-mail is spam

$$P(Y = 1 \,|\, x_1, \ldots, x_n) = \frac{P(Y = 1) \prod_{i=1}^{n} P(x_i \,|\, Y = 1)}{\sum_{y \in \{0,1\}} P(Y = y) \prod_{i=1}^{n} P(x_i \,|\, Y = y)}$$

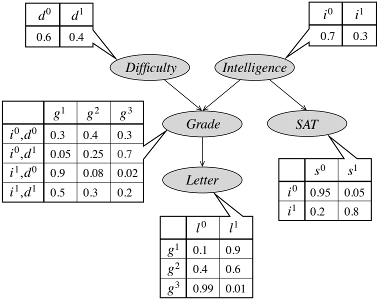
Example: Naive Bayes

$$P(Y, X_1, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



Example: Student

- Consider the following Student Bayesian network



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

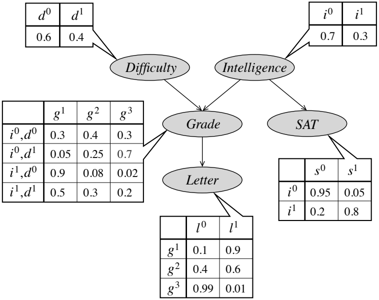
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- What is the joint distribution?

Example: Student

- Consider the following Student Bayesian network



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

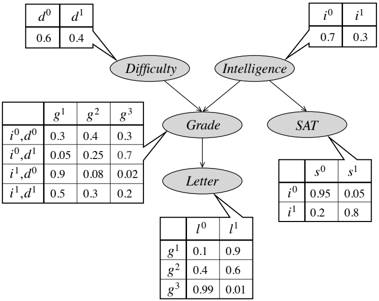
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- What is the joint distribution?

$$P(X_i, \ldots, X_n) = \prod_{i \in V} P(X_i \mid \mathrm{Pa}^G_{X_i})$$

- Consider the following STUDENT Bayesian network



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

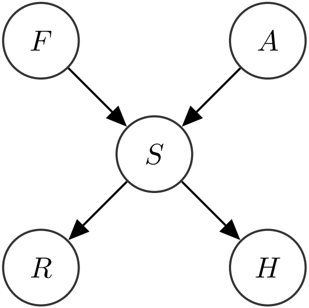
- What is the joint distribution?

$$P(X_i, \ldots, X_n) = \prod_{i \in V} P(X_i \mid \mathrm{Pa}^G_{X_i})$$

$$P(D, I, G, S, L) = P(D)P(I)P(G \mid I, D)P(S \mid I)P(L \mid G)$$

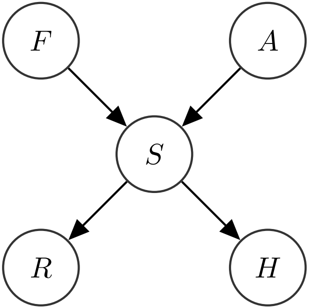
Example: Simple Medical Diagnosis

- The flu (F) causes sinus inflammation (S)
- Allergies (A) *also* cause sinus inflammation
- Sinus inflammation causes a runny nose (R)
- Sinus inflammation causes headaches (H)



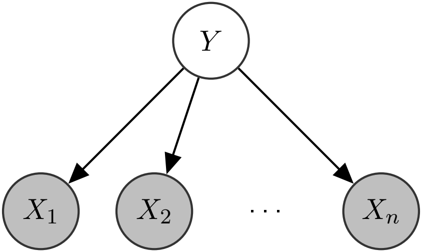
Example: Simple Medical Diagnosis

- The flu (F) causes sinus inflammation (S)
- Allergies (A) *also* cause sinus inflammation
- Sinus inflammation causes a runny nose (R)
- Sinus inflammation causes headaches (H)



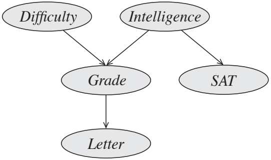
$$P(F, A, S, R, H) = P(F)P(A)P(S \mid F, A)P(R \mid S)P(H \mid S)$$

Bayesian Networks are Generative Models



- **Evidence** (observed variable) indicated by shaded node
- Can interpret Bayesian network as a **generative process**. For example, to *generate* an e-mail we
  1. Decide whether it is spam or not spam by sampling $y \sim P(Y)$
  2. For each word i, sample $x_i \sim P(X_i \mid Y = y)$

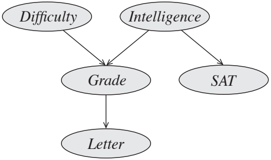
From Factorization to Independencies



- Joint distribution for above BN factors as

$$P(D, I, G, S, L) = P(D)P(I)P(G \mid I, D)P(S \mid I)P(L \mid G)$$

From Factorization to Independencies



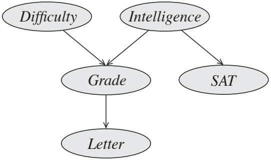
- Joint distribution for above BN factors as

$$P(D, I, G, S, L) = P(D)P(I)P(G \mid I, D)P(S \mid I)P(L \mid G)$$

- However, any distribution can be factored as (per chain rule)

$$P(D, I, G, S, L) = P(D)P(I \mid D)P(G \mid I, D)P(S \mid I, D, G) \\ \cdot P(L \mid I, D, G, S)$$

From Factorization to Independencies



- Joint distribution for above BN factors as

$$P(D, I, G, S, L) = P(D)P(I)P(G \mid I, D)P(S \mid I)P(L \mid G)$$

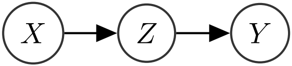
- However, any distribution can be factored as (per chain rule)

$$P(D, I, G, S, L) = P(D)P(I \mid D)P(G \mid I, D)P(S \mid I, D, G) \cdot P(L \mid I, D, G, S)$$

- Thus, if P factorizes over G, P exhibits the following independencies $(D \perp I)$ $(S \perp \{D, G\} \mid I)$ $(L \perp \{I, D, S\} \mid G)$ and others...

BN Structure Implies Conditional Independencies

- **Cascade** (Markov chain; causal trail; evidential trail; head-to-tail):



$(X \perp Y)$? $(X \perp Y \mid Z)$?

BN Structure Implies Conditional Independencies

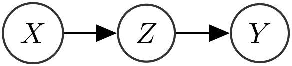
- **Cascade** (Markov chain; causal trail; evidential trail; head-to-tail):

$$\begin{tikzcd} X \arrow[r] & Z \arrow[r] & Y \end{tikzcd}$$

$(X \perp Y)$? No $\qquad$ $(X \perp Y \mid Z)$?

BN Structure Implies Conditional Independencies

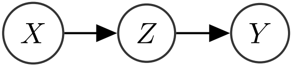
- **Cascade** (Markov chain; causal trail; evidential trail; head-to-tail):



$(X \perp Y)$? No $(X \perp Y \mid Z)$? Yes

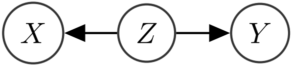
BN Structure Implies Conditional Independencies

- **Cascade** (Markov chain; causal trail; evidential trail; head-to-tail):



$(X \perp Y)$? No $(X \perp Y \mid Z)$? Yes

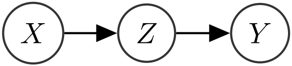
- **Common Cause** (tail-to-tail):



$(X \perp Y)$? $(X \perp Y \mid Z)$?

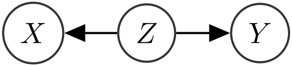
BN Structure Implies Conditional Independencies

- **Cascade** (Markov chain; causal trail; evidential trail; head-to-tail):



$(X \perp Y)$? No $(X \perp Y \mid Z)$? Yes

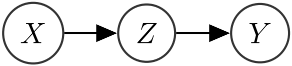
- **Common Cause** (tail-to-tail):



$(X \perp Y)$? No $(X \perp Y \mid Z)$?

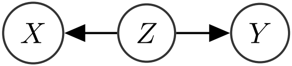
BN Structure Implies Conditional Independencies

- **Cascade** (Markov chain; causal trail; evidential trail; head-to-tail):



$(X \perp Y)$? No $\quad$ $(X \perp Y \mid Z)$? Yes

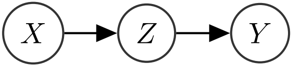
- **Common Cause** (tail-to-tail):



$(X \perp Y)$? No $\quad$ $(X \perp Y \mid Z)$? Yes (Naive Bayes)

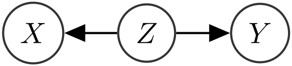
BN Structure Implies Conditional Independencies

- **Cascade** (Markov chain; causal trail; evidential trail; head-to-tail):



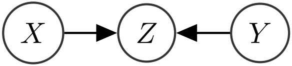
$(X \perp Y)$? No $\quad\quad$ $(X \perp Y \mid Z)$? Yes

- **Common Cause** (tail-to-tail):



$(X \perp Y)$? No $\quad\quad$ $(X \perp Y \mid Z)$? Yes (Naive Bayes)

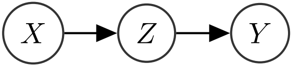
- **Common Effect** (v-structure, "explaining away"; head-to-head):



$(X \perp Y)$? $\quad\quad$ $(X \perp Y \mid Z)$?

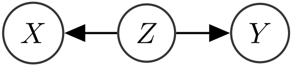
BN Structure Implies Conditional Independencies

- **Cascade** (Markov chain; causal trail; evidential trail; head-to-tail):



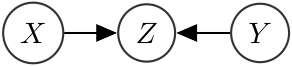
$(X \perp Y)$? No $(X \perp Y \mid Z)$? Yes

- **Common Cause** (tail-to-tail):



$(X \perp Y)$? No $(X \perp Y \mid Z)$? Yes (Naive Bayes)

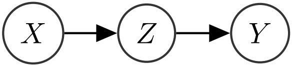
- **Common Effect** (v-structure, "explaining away"; head-to-head):



$(X \perp Y)$? Yes $(X \perp Y \mid Z)$?

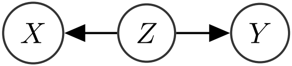
BN Structure Implies Conditional Independencies

- **Cascade** (Markov chain; causal trail; evidential trail; head-to-tail):



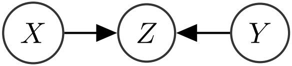
$(X \perp Y)$? No $\qquad$ $(X \perp Y \mid Z)$? Yes

- **Common Cause** (tail-to-tail):



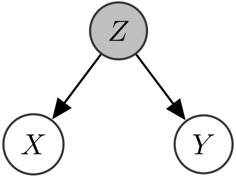
$(X \perp Y)$? No $\qquad$ $(X \perp Y \mid Z)$? Yes (Naive Bayes)

- **Common Effect** (v-structure, "explaining away"; head-to-head):



$(X \perp Y)$? Yes $\qquad$ $(X \perp Y \mid Z)$? No

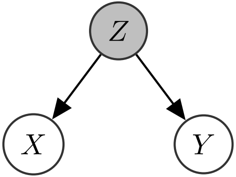
Common Cause



We will show that $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$ for any distribution $P(X, Y, Z)$ that factors according to this graph, i.e.,

$$P(X, Y, Z) = P(Z)P(X \mid Z)P(Y \mid Z)$$

Common Cause



We will show that $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$ for any distribution $P(X, Y, Z)$ that factors according to this graph, i.e.,

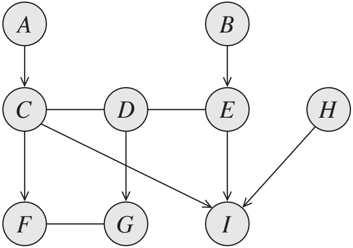
$$P(X, Y, Z) = P(Z)P(X \mid Z)P(Y \mid Z)$$

Proof

$$P(X, Y \mid Z) = \frac{P(X, Y, Z)}{P(Z)} = P(X \mid Z)P(Y \mid Z)$$

Active Trail

Let G be a BN structure and $X_1 \leftrightharpoons \ldots \leftrightharpoons X_n$ be a *trail* in G



Active Trail

Let G be a BN structure and $X_1 \leftrightharpoons \ldots \leftrightharpoons X_n$ be a *trail* in G

Let $\boldsymbol{Z}$ be a subset of observed variables

Active Trail

Let G be a BN structure and $X_1 \leftrightharpoons \ldots \leftrightharpoons X_n$ be a *trail* in G

Let $\boldsymbol{Z}$ be a subset of observed variables

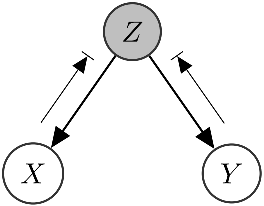
The trail is **active** (i.e., dependency/information flow) given $\boldsymbol{Z}$ if

- For every v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, X_i or one of its descendents is in $\boldsymbol{Z}$
- No other node along the trail is in $\boldsymbol{Z}$

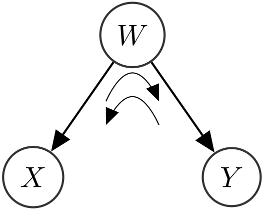
D-Separation ("Directed Separation")

Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be three sets of nodes in graph G

- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **d-separated** given $\boldsymbol{Z}$ (d-sep$_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no "active trail" between any node $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ given $\boldsymbol{Z}$
- Alternatively, conditioning on $\boldsymbol{Z}$ "blocks" the path from $\boldsymbol{X}$ to $\boldsymbol{Y}$
- If d-sep$_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$, then $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ (soundness)



No active trail



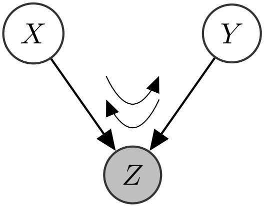
Active trail ($W \notin Z$)

(First proposed by Pearl in 1986)

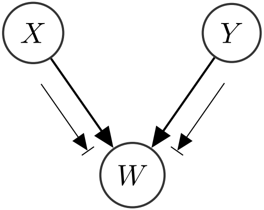
D-Separation ("Directed Separation")

Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be three sets of nodes in graph G

- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **d-separated** given $\boldsymbol{Z}$ (d-sep$_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no "active trail" between any node $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ given $\boldsymbol{Z}$
- Alternatively, conditioning on $\boldsymbol{Z}$ "blocks" the path from $\boldsymbol{X}$ to $\boldsymbol{Y}$
- If d-sep$_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$, then $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ (soundness)



Active trail



No active trail ($W \notin Z$)

D-Separation ("Directed Separation")

Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be three sets of nodes in graph G

- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **d-separated** given $\boldsymbol{Z}$ ($\text{d-sep}_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no "active trail" between any node $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ given $\boldsymbol{Z}$

D-Separation ("Directed Separation")

Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be three sets of nodes in graph G

- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **d-separated** given $\boldsymbol{Z}$ ($\text{d-sep}_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no "active trail" between any node $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ given $\boldsymbol{Z}$
- For a BN structure G, we define the **global Markov independencies** as the set of independencies that correspond to d-separation

$$I(G) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} : \text{d-sep}_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}))\}$$

D-Separation ("Directed Separation")

Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be three sets of nodes in graph G

- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **d-separated** given $\boldsymbol{Z}$ ($\text{d-sep}_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no "active trail" between any node $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ given $\boldsymbol{Z}$
- For a BN structure G, we define the **global Markov independencies** as the set of independencies that correspond to d-separation

$$I(G) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} : \text{d-sep}_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}))\}$$

- D-separation reduces reasoning over statistical independencies (hard problem) to analyzing connectivity in graphs (easy problem)

D-Separation ("Directed Separation")

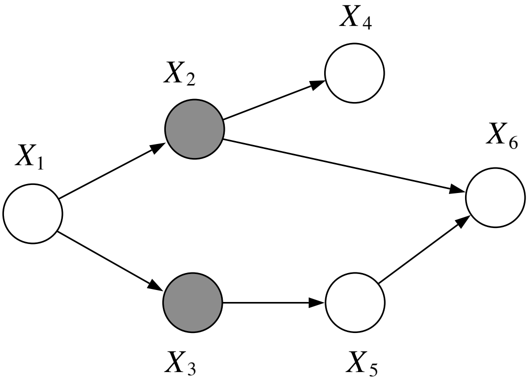
Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be three sets of nodes in graph G

- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **d-separated** given $\boldsymbol{Z}$ ($\text{d-sep}_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no "active trail" between any node $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ given $\boldsymbol{Z}$
- For a BN structure G, we define the **global Markov independencies** as the set of independencies that correspond to d-separation

$$I(G) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} : \text{d-sep}_G(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}))\}$$

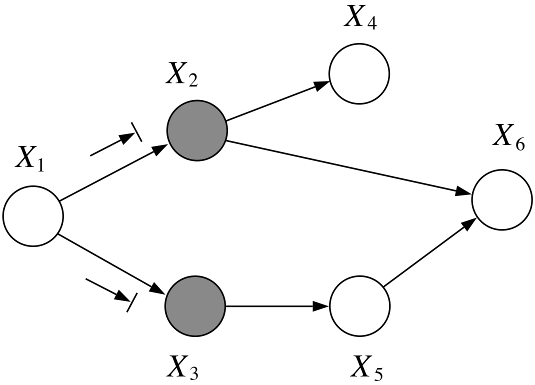
- D-separation reduces reasoning over statistical independencies (hard problem) to analyzing connectivity in graphs (easy problem)
- Enables us to reduce the Bayesian network to only the variables relevant to answering a query

D-Separation: Example



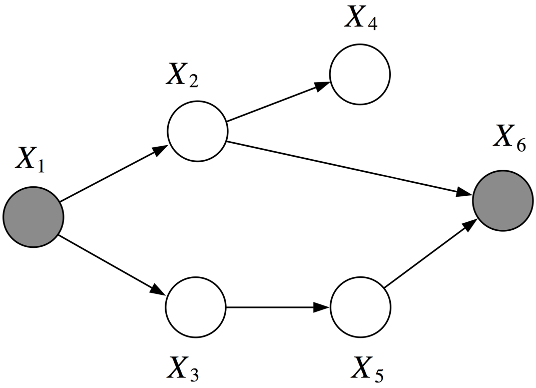
Are X_1 and X_5 d-separated given X_2 and X_3?

D-Separation: Example



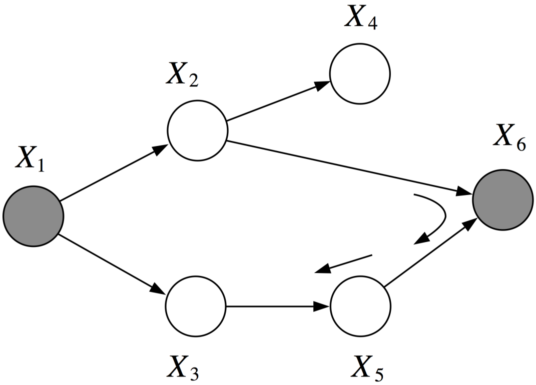
Are X_1 and X_5 d-separated given X_2 and X_3? Yes

D-Separation: Example



Are X_2 and X_3 d-separated given X_1 and X_6?

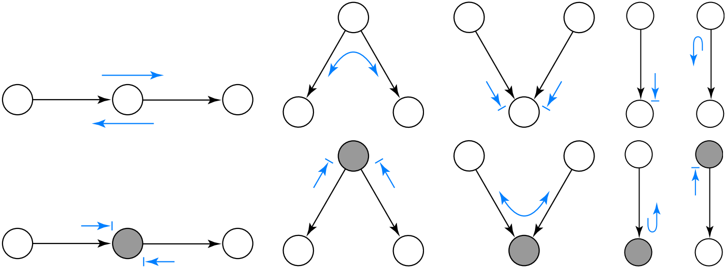
D-Separation: Example



Are X_2 and X_3 d-separated given X_1 and X_6? No (v-structure)

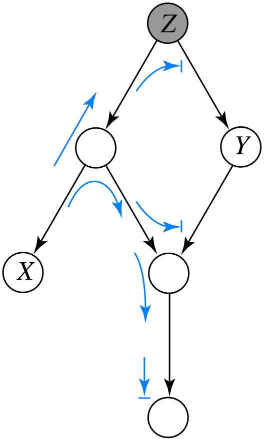
Bayes Ball Algorithm (due to Ross Shachter)

- An alternative algorithm for identifying active trails and d-separation
- An undirected path is active iff a Bayes ball travelling along it never encounters a "stop" symbol

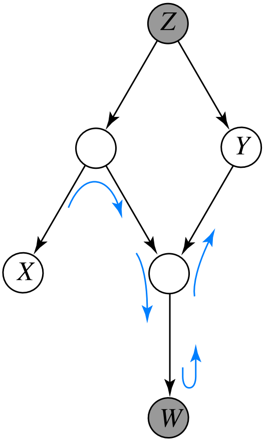


- No active paths from X to Y when $\boldsymbol{Z}$ are shaded $\longrightarrow (X \perp Y \mid \boldsymbol{Z})$

D-Separation: Example



No active paths
$(X \perp Y \mid Z)$



One active path
$(X \not\perp Y \mid W, Z)$

D-Separation Algorithm

Given BN structure G, determine whether X and Y d-separated given $\boldsymbol{Z}$

1. Traverse graph from leaves to root (bottom-up) and mark any node that is in $\boldsymbol{Z}$ or has a descendant in $\boldsymbol{Z}$ (i.e., v-structures)
2. Perform breadth-first search from X along active trails (i.e., stopping at nodes in $\boldsymbol{Z}$ or marked nodes in the middle of a v-structure) generating *reachable set* $\boldsymbol{R}$
   - Requires bookkeeping to keep track of whether node was reached via children or parents
3. X and Y are d-separated iff $Y \notin \boldsymbol{R}$

Try this with the graphs on the previous slide

D-Separation: Soundness & Completeness

For a BN structure G and any $X, Y, \boldsymbol{Z}$, we would like

$$\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Leftrightarrow P \vDash (X \perp Y \mid \boldsymbol{Z})$$

D-Separation: Soundness & Completeness

For a BN structure G and any $X, Y, \boldsymbol{Z}$, we would like

$$\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Leftrightarrow P \models (X \perp Y \mid \boldsymbol{Z})$$

Definition (Soundness)

If P factorizes according to G, then $\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Rightarrow P \models (X \perp Y \mid \boldsymbol{Z})$

D-Separation: Soundness & Completeness

For a BN structure G and any $X, Y, \boldsymbol{Z}$, we would like

$$\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Leftrightarrow P \vDash (X \perp Y \mid \boldsymbol{Z})$$

Definition (Soundness)

If P factorizes according to G, then $\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Rightarrow P \vDash (X \perp Y \mid \boldsymbol{Z})$

Definition (Completeness)

For any P that factorizes per G, $P \vDash (X \perp Y \mid \boldsymbol{Z}) \Rightarrow \text{d-sep}_G(X, Y \mid \boldsymbol{Z})$

D-Separation: Soundness & Completeness

For a BN structure G and any $X, Y, \boldsymbol{Z}$, we would like

$$\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Leftrightarrow P \models (X \perp Y \mid \boldsymbol{Z})$$

Definition (Soundness)

If P factorizes according to G, then $\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Rightarrow P \models (X \perp Y \mid \boldsymbol{Z})$

Definition (Completeness)

For any P that factorizes per G, $P \models (X \perp Y \mid \boldsymbol{Z}) \Rightarrow \text{d-sep}_G(X, Y \mid \boldsymbol{Z})$

Does "completeness" imply the contrapositive: If X and Y are *not* d-separated given $\boldsymbol{Z}$, then $P \not\models (X \perp Y \mid \boldsymbol{Z})$ for all P that factorize per G?

D-Separation: Soundness & Completeness

For a BN structure G and any $X, Y, \boldsymbol{Z}$, we would like

$$\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Leftrightarrow P \models (X \perp Y \mid \boldsymbol{Z})$$

Definition (Soundness)

If P factorizes according to G, then $\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Rightarrow P \models (X \perp Y \mid \boldsymbol{Z})$

Definition (Completeness)

For any P that factorizes per G, $P \models (X \perp Y \mid \boldsymbol{Z}) \Rightarrow \text{d-sep}_G(X, Y \mid \boldsymbol{Z})$

Does "completeness" imply the contrapositive: If X and Y are *not* d-separated given $\boldsymbol{Z}$, then $P \not\models (X \perp Y \mid \boldsymbol{Z})$ for all P that factorize per G?
No! G specifies the topology, not the parameters

D-Separation: Soundness & Completeness

Consider the following Bayesian network, where X, Y, Z are boolean

$$Z \rightarrow X \rightarrow Y$$

$$P(Z) = 0.9$$

$$P(X \mid Z) = 1 \qquad P(X \mid \neg Z) = 1$$

$$P(Y \mid X) = 0.5 \qquad P(Y \mid \neg X) = 0.5$$

D-Separation: Soundness & Completeness

For a BN structure G and any $X, Y, \boldsymbol{Z}$, we would like

$$\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Leftrightarrow P \models (X \perp Y \mid \boldsymbol{Z})$$

Definition (Soundness)

If P factorizes according to G, then $\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Rightarrow P \models (X \perp Y \mid \boldsymbol{Z})$

Definition (Completeness (alternative))

If $P \models (X \perp Y \mid \boldsymbol{Z})$ for *all* distributions P that factorize over G, then $\text{d-sep}_G(X, Y \mid \boldsymbol{Z})$

D-Separation: Soundness & Completeness

For a BN structure G and any $X, Y, \boldsymbol{Z}$, we would like

$$\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Leftrightarrow P \models (X \perp Y \mid \boldsymbol{Z})$$

Definition (Soundness)

If P factorizes according to G, then $\text{d-sep}_G(X, Y \mid \boldsymbol{Z}) \Rightarrow P \models (X \perp Y \mid \boldsymbol{Z})$

Definition (Completeness (alternative))

If $P \models (X \perp Y \mid \boldsymbol{Z})$ for *all* distributions P that factorize over G, then $\text{d-sep}_G(X, Y \mid \boldsymbol{Z})$

Theorem

Let G be a BN structure. If X and Y are **not d-separated** *given $\boldsymbol{Z}$ in G, then X and Y are* **dependent given** *$\boldsymbol{Z}$ in some distribution P that factorizes over G*

D-Separation: Soundness & Completeness

Theorem (Meek 1995)

For almost all distributions P that factorize over G (except for a set of measure zero), we have $I(P) = I(G)$

In other words, the set of parameterizations for which the distribution is unfaithful are of measure zero.

Implies that most distributions that factorize over G are faithful.

Independence Maps

- Let $I(P) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})\}$ be the set of independence assertions that hold in P (i.e., $P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$)
- A BN structure G is an **I-map** (independence map) for a set of independencies I if $I(G) \subseteq I$
- A BN structure G is an **I-map** for P if G is an I-map for $I(P)$, i.e., $I(G) \subseteq I(P)$

Independence Maps

- Let $I(P) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})\}$ be the set of independence assertions that hold in P (i.e., $P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$)
- A BN structure G is an **I-map** (independence map) for a set of independencies I if $I(G) \subseteq I$
- A BN structure G is an **I-map** for P if G is an I-map for $I(P)$, i.e., $I(G) \subseteq I(P)$
  - Any independence asserted by G must hold in P

Independence Maps

- Let $I(P) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})\}$ be the set of independence assertions that hold in P (i.e., $P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$)
- A BN structure G is an **I-map** (independence map) for a set of independencies I if $I(G) \subseteq I$
- A BN structure G is an **I-map** for P if G is an I-map for $I(P)$, i.e., $I(G) \subseteq I(P)$
  - Any independence asserted by G must hold in P
  - Converse need not be true—P may have additional independencies not reflected in G

Independence Maps

- Let $I(P) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})\}$ be the set of independence assertions that hold in P (i.e., $P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$)
- A BN structure G is an **I-map** (independence map) for a set of independencies I if $I(G) \subseteq I$
- A BN structure G is an **I-map** for P if G is an I-map for $I(P)$, i.e., $I(G) \subseteq I(P)$
  - Any independence asserted by G must hold in P
  - Converse need not be true—P may have additional independencies not reflected in G
  - Trivial case: A fully connected graph G is an I-map for *any* distribution since $I(G) = \emptyset \subseteq I(P)\ \forall P$

Representation Theorem

Theorem (Verma & Pearl, 1998)

Given a BN structure G and joint distribution P over a set of random variables, P factorizes over G **iff** *G is an I-map for P*

Representation Theorem

Theorem (Verma & Pearl, 1998)

Given a BN structure G and joint distribution P over a set of random variables, P factorizes over G **iff** *G is an I-map for P*

- If $I(G) \subseteq I(P)$, *any* conditional independency expressed by G holds *for all* distributions P that factorize over G
- If $I(G) \subseteq I(P)$, *any* any conditional dependency expressed by G holds *for some* distributions that factorize over G

Representation Theorem: Proof

Consider one direction: P factorizes over $G \Leftarrow G$ is an I-map for P

Proof

Let T be a topological ordering of the nodes in G, i.e., $\forall i \in T$, Pa$_i^G$ appear before i

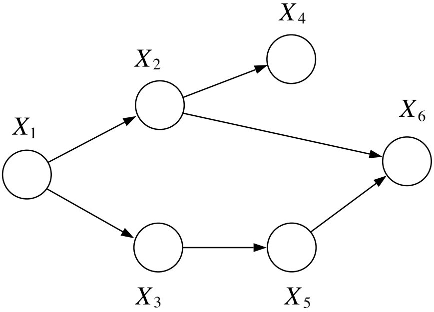
Let ν_i be the set of nodes appearing before i in T, excluding Pa$_i^G$

From $I_l(G)$, we have that $\{X_i \perp X_{\nu_i} \mid \mathit{Pa}_{X_i}^G\}$

Since $I(G) \subseteq I(P)$,

$$P(X_1, \ldots, X_n) = \prod_{i \in T} P(X_i \mid X_{\nu_i}, \mathit{Pa}_{X_i}^G) = \prod_{i \in T} P(X_i \mid \mathit{Pa}_{X_i}^G)$$

Representation Theorem: Proof (Example)



$T = \{1, 2, 3, 4, 5, 6\}$

$\nu_1 = \emptyset,\ \nu_2 = \emptyset,\ \nu_3 = \{2\},\ \nu_4 = \{1, 3\}\ \nu_5 = \{1, 2, 4\},\ \nu_6 = \{1, 3, 4\}$

$$P(_1, \ldots, X_6) = \prod_{i \in T} P(X_i \mid X_{\nu_i}, \mathsf{Pa}^G_{X_i})$$

$$= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_4 \mid X_2)P(X_5 \mid X_3)P(X_6 \mid X_2, X_5)$$

I-Equivalence

- Different BN structures are **I-equivalent** if they encode the same conditional independencies (and, in turn, the same distributions)
- For a given P, any equivalent BN structure is equally valid

Definition (Skeleton)

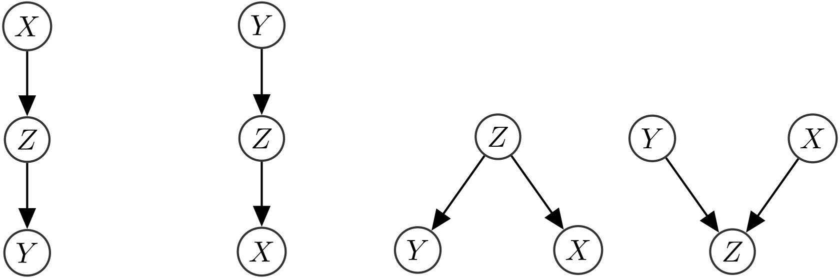
The **skeleton** of a Bayesian network graph G over $\mathcal{X}$ is an undirected graph over $\mathcal{X}$ with an undirected edge $\{X, Y\}$ for every edge (X, Y) in G

Theorem

Let G_1 and G_2 be two graphs over $\mathcal{X}$. If G_1 and G_2 have the same skeleton and the same set of v-structures, then they are I-equivalent

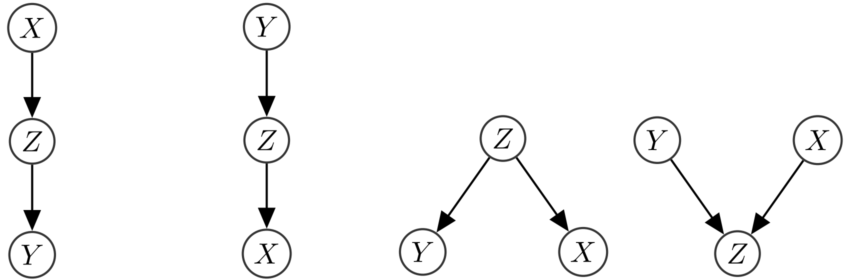
I-Equivalence

Which of the following are equivalent?



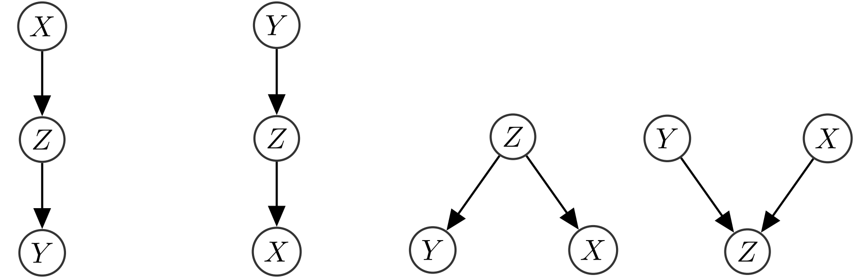
I-Equivalence

Which of the following are equivalent?



I-Equivalence

Which of the following are equivalent?



$(X \perp Y \mid Z)$ $(Y \perp X \mid Z)$ $(X \perp Y \mid Z)$ $(X \not\perp Y \mid Z)$

Distributions to Graphs

- If $I(G) \subseteq I(P)$, G is an I-map for P and we can use G to identify (and exploit) independencies in P
- Is G missing independencies?
- A graph G is a **minimal I-map** for a set of independencies I if $I(G) \subseteq I$ and removing a single edge from G results in $I(\bar{G}) \not\subseteq I$

Distributions to Graphs

Given a distribution P and its independencies $I(P)$, how do we generate the minimal I-map? (Hint: Recall the factorization proof)

Algorithm 3.2 Procedure to build a minimal I-map given an ordering

---

Procedure Build-Minimal-I-Map (
  $X_1, \ldots, X_n$ // an ordering of random variables in $\mathcal{X}$
  $\mathcal{I}$ // Set of independencies
)
  Set $\mathcal{G}$ to an empty graph over $\mathcal{X}$
  for $i = 1, \ldots, n$
    $\boldsymbol{U} \leftarrow \{X_1, \ldots, X_{i-1}\}$ // $\boldsymbol{U}$ is the current candidate for parents of X_i
    for $\boldsymbol{U}' \subseteq \{X_1, \ldots, X_{i-1}\}$
      if $\boldsymbol{U}' \subset \boldsymbol{U}$ and $(X_i \perp \{X_1, \ldots, X_{i-1}\} - \boldsymbol{U}' \mid \boldsymbol{U}') \in \mathcal{I}$ **then**
        $\boldsymbol{U} \leftarrow \boldsymbol{U}'$
      // At this stage $\boldsymbol{U}$ is a minimal set satisfying $(X_i \perp \{X_1, \ldots, X_{i-1}\} - \boldsymbol{U} \mid \boldsymbol{U})$
      // Now set $\boldsymbol{U}$ to be the parents of X_i
    for $X_j \in \boldsymbol{U}$
      Add $X_j \rightarrow X_i$ to $\mathcal{G}$
  return $\mathcal{G}$

Perfect Maps

- $I(G) \subseteq I(P)$ if G is an I-map for P, but is G missing independencies?
- A graph G is a **minimal I-map** for a set of independencies I if $I(G) \subseteq I$ and removing a single edge from G results in $I(\bar{G}) \not\subseteq I$
- A graph G is a **perfect map (P-map)** for P if $I(G) = I(P)$

If $I(G) = I(P)$, then we can read independencies of P directly from G

Perfect Maps

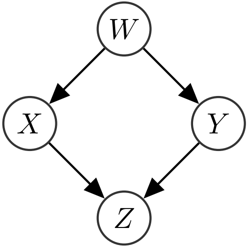
Not all distributions P have a perfect map

Let P be a distribution over $\mathcal{X} = \{W, X, Y, Z\}$ such that
$P \models = \{(W \perp Z \mid X, Y), (X \perp Y \mid W, Z)\}$

Perfect Maps

Not all distributions P have a perfect map

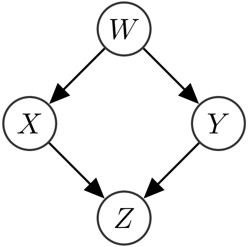
Let P be a distribution over $\mathcal{X} = \{W, X, Y, Z\}$ such that
$P \models= \{(W \perp Z \mid X, Y), (X \perp Y \mid W, Z)\}$



Perfect Maps

Not all distributions P have a perfect map

Let P be a distribution over $\mathcal{X} = \{W, X, Y, Z\}$ such that $P \models= \{(W \perp Z \mid X, Y), (X \perp Y \mid W, Z)\}$

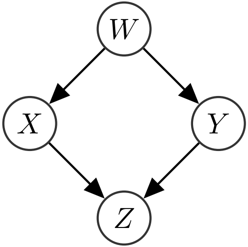


$(W \perp Z \mid X, Y)$

Perfect Maps

Not all distributions P have a perfect map

Let P be a distribution over $\mathcal{X} = \{W, X, Y, Z\}$ such that
$P \models = \{(W \perp Z \mid X, Y), (X \perp Y \mid W, Z)\}$

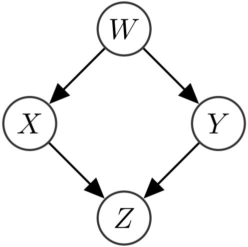


$(W \perp Z \mid X, Y)$
$(X \not\perp Y \mid W, Z)$

Perfect Maps

Not all distributions P have a perfect map

Let P be a distribution over $\mathcal{X} = \{W, X, Y, Z\}$ such that
$P \models = \{(W \perp Z \mid X, Y), (X \perp Y \mid W, Z)\}$

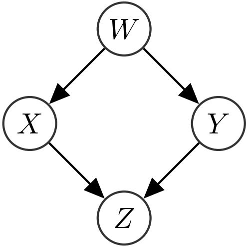


$(W \perp Z \mid X, Y)$
$(X \not\perp Y \mid W, Z)$
$(X \perp Y \mid W)$

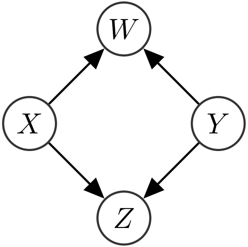
Perfect Maps

Not all distributions P have a perfect map

Let P be a distribution over $\mathcal{X} = \{W, X, Y, Z\}$ such that
$P \models= \{(W \perp Z \mid X, Y), (X \perp Y \mid W, Z)\}$

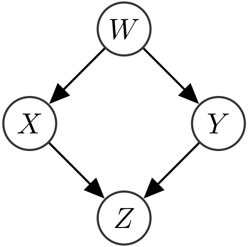


$(W \perp Z \mid X, Y)$
$(X \not\perp Y \mid W, Z)$
$(X \perp Y \mid W)$

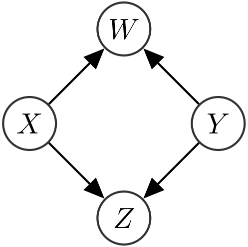


Not all distributions P have a perfect map

Let P be a distribution over $\mathcal{X} = \{W, X, Y, Z\}$ such that
$P \models = \{(W \perp Z \mid X, Y), (X \perp Y \mid W, Z)\}$



$(W \perp Z \mid X, Y)$
$(X \not\perp Y \mid W, Z)$
$(X \perp Y \mid W)$

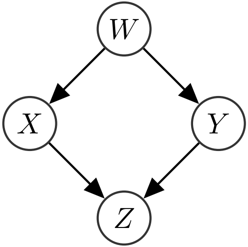


$(W \perp Z \mid X, Y)$

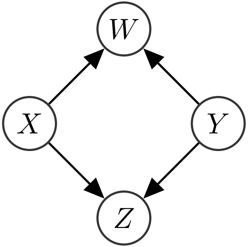
Perfect Maps

Not all distributions P have a perfect map

Let P be a distribution over $\mathcal{X} = \{W, X, Y, Z\}$ such that
$P \models = \{(W \perp Z \mid X, Y), (X \perp Y \mid W, Z)\}$



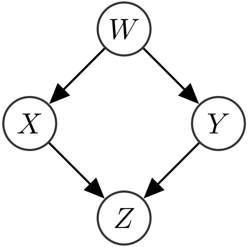
$(W \perp Z \mid X, Y)$
$(X \not\perp Y \mid W, Z)$
$(X \perp Y \mid W)$



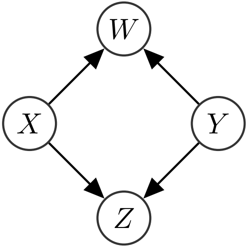
$(W \perp Z \mid X, Y)$
$(X \not\perp Y \mid W, Z)$

Not all distributions P have a perfect map

Let P be a distribution over $\mathcal{X} = \{W, X, Y, Z\}$ such that
$P \models \{(W \perp Z \mid X, Y), (X \perp Y \mid W, Z)\}$



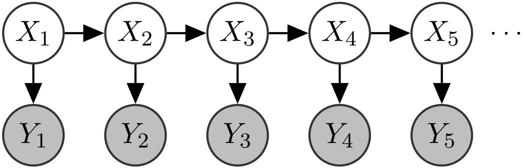
$(W \perp Z \mid X, Y)$
$(X \not\perp Y \mid W, Z)$
$(X \perp Y \mid W)$



$(W \perp Z \mid X, Y)$
$(X \not\perp Y \mid W, Z)$
$(X \perp Y)$

What are some frequently used Bayesian network models?

Hidden Markov Models (HMMs)

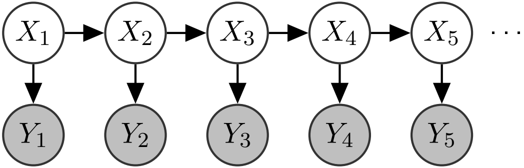


- Commonly used to model speech recognition and NLP problems
- Joint distribution can be factored as

$$P(\boldsymbol{X}, \boldsymbol{Y}) = P(X_1)P(Y_1 \mid X_1) \prod_{t=2}^{T} P(X_t \mid X_{t-1}) P(Y_t \mid X_t)$$

  - $P(X_1)$ is the distribution over the starting state
  - $P(X_t \mid X_{t-1})$ is the **transition** probability
  - $P(Y_t \mid X_t)$ is the **emission** (observation) probability

Hidden Markov Models (HMMs)

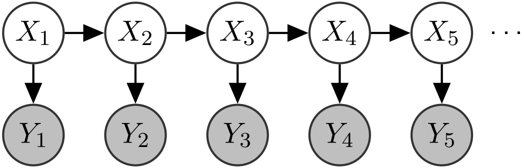


- Joint distribution can be factored as

$$P(\boldsymbol{X}, \boldsymbol{Y}) = P(X_1)P(Y_1 \mid X_1) \prod_{t=2}^{T} P(X_t \mid X_{t-1}) P(Y_t \mid X_t)$$

- A **homogeneous** HMM uses the same parameters (α and β) for the transition and emission distributions (aka parameter sharing)
  - $P(X_t \mid X_{t-1}) = \beta_{X_t, X_{t-1}}$, $P(Y_t \mid X_t) = \alpha_{Y_t, X_t}$
- How many parameters are needed?

Hidden Markov Models (HMMs)



- Joint distribution can be factored as

$$P(\boldsymbol{X}, \boldsymbol{Y}) = P(X_1)P(Y_1 \mid X_1) \prod_{t=2}^{T} P(X_t \mid X_{t-1}) P(Y_t \mid X_t)$$

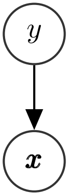
- A **homogeneous** HMM uses the same parameters (α and β) for the transition and emission distributions (aka parameter sharing)
  - $P(X_t \mid X_{t-1}) = \beta_{X_t, X_{t-1}}$, $P(Y_t \mid X_t) = \alpha_{Y_t, X_t}$
- How many parameters are needed? $(|X_i| - 1)|X_i| + (|Y_i| - 1)|X_i|$ (e.g., $2 + 2 = 4$ if Y_i and X_i are binary)

Mixture of Gaussians

- Consider an n-dim multivariate Gaussian $x \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Mixture of Gaussians

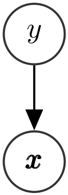


- Consider an n-dim multivariate Gaussian $x \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- Consider k different Gaussians $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ and let $y \in \{1, \ldots, k\}$ be an index with distribution $p(y)$ (alt θ)

Mixture of Gaussians



- Consider an n-dim multivariate Gaussian $x \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

- Consider k different Gaussians $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ and let $y \in \{1, \ldots, k\}$ be an index with distribution $p(y)$ (alt θ)
- Mixture of Gaussians distribution $p(y, \boldsymbol{x})$ can be sampled as
  - Sample $y \sim p(y)$ (sample which Gaussian)
  - Sample $x \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$

Mixture of Gaussians

- The marginal distribution $p(x) = \sum_{y \in \{1,\ldots,k\}} p(y, x)$