Probabilistic Graphical Models

Bayesian Networks
Markov Random Fields

Luigi Portinale

University of Piemonte Orientale, Italy

March 7-10, 2017

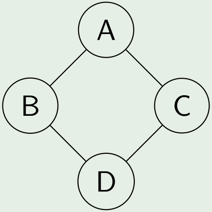
Probabilistic Graphical Models

Probabilistic Graphical Models (PGM)

A probabilistic graphical model is a pair $\langle G, \mathcal{P} \rangle$ such that G is a graph whose nodes correspond to (discrete) random variables and edges to dependency relations, while $\mathcal{P}$ is a probability distribution over the variables corresponding to nodes in G.

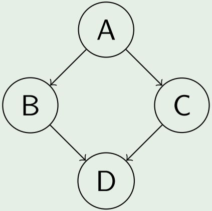
- if the graph G is undirected we have un Undirected Graphical Model (UGM) also called Markov Random Field (MRF) or Markov Network (MN);
- if the graph G is a DAG we have a Bayesian Network (BN) also called Belief Network

Markov Random Field



Edge between X and Y states that X and Y depend from each other

Bayesian Network



Edge from X to Y states that X influences Y and that the influence has a directionality (e.g. causality)

Bayesian Networks

Bayesian Network: definition

A **Bayesian Network** is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is a DAG whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i \rightarrow X_j) \in E$ represents a direct influence of X_i over X_j (e.g. X_i "causes" X_j or "the presence of X_i suggests the presence of X_j");
- $\mathcal{P}$ is a probability distribution over the variables represented by V, such that

$$\mathcal{P}(X_1, \ldots X_n) = \prod_{i=1}^{n} \mathcal{P}(X_i | \pi(X_i))$$

with $\pi(X) = \{Y \in V : Y \text{ is a parent of } X \text{ in } G\}$

We say that $\mathcal{P}$ **factorizes** over G.

Bayesian Network: definition

A Bayesian Network is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is a DAG whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i \rightarrow X_j) \in E$ represents a direct influence of X_i over X_j (e.g. X_i "causes" X_j or "the presence of X_i suggests the presence of X_j");
- $\mathcal{P}$ is a probability distribution over the variables represented by V, such that

$$\mathcal{P}(X_1, \ldots X_n) = \prod_{i=1}^{n} \mathcal{P}(X_i | \pi(X_i))$$

with $\pi(X) = \{Y \in V : Y \text{ is a parent of } X \text{ in } G\}$

We say that $\mathcal{P}$ factorizes over G.

Bayesian Network: definition

A Bayesian Network is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is a DAG whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i \to X_j) \in E$ represents a direct influence of X_i over X_j (e.g. X_i "causes" X_j or "the presence of X_i suggests the presence of X_j");
- $\mathcal{P}$ is a probability distribution over the variables represented by V, such that

$$\mathcal{P}(X_1, \ldots X_n) = \prod_{i=1}^{n} \mathcal{P}(X_i | \pi(X_i))$$

with $\pi(X) = \{Y \in V : Y \text{ is a parent of } X \text{ in } G\}$

We say that $\mathcal{P}$ factorizes over G.

Bayesian Network: definition

A Bayesian Network is a pair $\langle G, \mathcal{P} \rangle$ where:

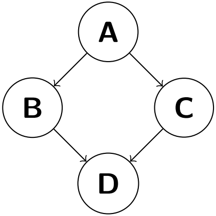
- $G = \langle V, E \rangle$ is a DAG whose vertices $V = \{X_1, \dots X_n\}$ represent (discrete) random variables and and edge $(X_i \rightarrow X_j) \in E$ represents a direct influence of X_i over X_j (e.g. X_i "causes" X_j or "the presence of X_i suggests the presence of X_j");
- $\mathcal{P}$ is a probability distribution over the variables represented by V, such that

$$\mathcal{P}(X_1, \dots X_n) = \prod_{i=1}^{n} \mathcal{P}(X_i | \pi(X_i))$$

with $\pi(X) = \{Y \in V : Y \text{ is a parent of } X \text{ in } G\}$

We say that $\mathcal{P}$ factorizes over G.

Bayesian Network: an example



$$P(A,B,C,D) = P(A)P(B|A)P(C|A)P(D|B,C)$$

A = 0	A = 1
0.3	0.7

$P(A)$

B \ A	0	1
0	0.2	0.65
1	0.8	0.35

$P(B|A)$

C \ A	0	1
0	0.1	0.25
1	0.9	0.75

$P(C|A)$

D \ B C	00	01	10	11
0	0.7	0.4	0.5	0.9
1	0.3	0.6	0.5	0.1

$P(D|B,C)$

BN Joint Distribution

(A,B,C,D)	$P(A)$	$P(B\|A)$	$P(C\|A)$	$P(D\|BC)$	$\mathcal{P}(A,B,C,D)$
$(0,0,0,0)$	0.3	0.2	0.1	0.7	0.0042
$(0,0,0,1)$	0.3	0.2	0.1	0.3	0.0018
$(0,0,1,0)$	0.3	0.2	0.9	0.4	0.0216
$(0,0,1,1)$	0.3	0.2	0.9	0.6	0.0324
$(0,1,0,0)$	0.3	0.8	0.1	0.5	0.0120
$(0,1,0,1)$	0.3	0.8	0.1	0.5	0.0120
$(0,1,1,0)$	0.3	0.8	0.9	0.9	0.1944
$(0,1,1,1)$	0.3	0.8	0.9	0.1	0.0216
$(1,0,0,0)$	0.7	0.65	0.25	0.7	0.0796
$(1,0,0,1)$	0.7	0.65	0.25	0.3	0.0341
$(1,0,1,0)$	0.7	0.65	0.75	0.4	0.1365
$(1,0,1,1)$	0.7	0.65	0.75	0.6	0.2048
$(1,1,0,0)$	0.7	0.35	0.25	0.5	0.0306
$(1,1,0,1)$	0.7	0.35	0.25	0.5	0.0306
$(1,1,1,0)$	0.7	0.35	0.75	0.9	0.1654
$(1,1,1,1)$	0.7	0.35	0.75	0.1	0.0184

Markov Random Fields

Pairwise Markov Random Field: definition

A Pairwise Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each edge $(X_i - X_j)$ is associated with a factor or potential $\Phi_{i,j} : \mathcal{D}(X_i) \times \mathcal{D}(X_j) \to \mathbb{R}^+ \cup \{0\}$ ($\mathcal{D}(X)$ domain of X)
- $\mathcal{P}(X_1, \ldots X_n) = \frac{1}{Z} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$

- $Z = \sum_{X_1,\ldots X_n} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

Pairwise Markov Random Field: definition

A Pairwise Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each edge $(X_i - X_j)$ is associated with a factor or potential $\Phi_{i,j} : \mathcal{D}(X_i) \times \mathcal{D}(X_j) \to \mathbb{R}^+ \cup \{0\}$ ($\mathcal{D}(X)$ domain of X)
- $\mathcal{P}(X_1, \ldots X_n) = \frac{1}{Z} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$

- $Z = \sum_{X_1, \ldots X_n} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

Pairwise Markov Random Field: definition

A Pairwise Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each edge $(X_i - X_j)$ is associated with a factor or potential $\Phi_{i,j} : \mathcal{D}(X_i) \times \mathcal{D}(X_j) \to \mathbb{R}^+ \cup \{0\}$ ($\mathcal{D}(X)$ domain of X)
- $\mathcal{P}(X_1, \ldots X_n) = \frac{1}{Z} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$

- $Z = \sum_{X_1,\ldots X_n} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

Pairwise Markov Random Field: definition

A Pairwise Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \dots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each edge $(X_i - X_j)$ is associated with a factor or potential $\Phi_{i,j} : \mathcal{D}(X_i) \times \mathcal{D}(X_j) \to \mathbb{R}^+ \cup \{0\}$ ($\mathcal{D}(X)$ domain of X)
- $\mathcal{P}(X_1, \dots X_n) = \frac{1}{Z} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$

- $Z = \sum_{X_1,\dots X_n} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

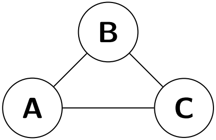
Pairwise Markov Random Field: definition

A Pairwise Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each edge $(X_i - X_j)$ is associated with a factor or potential $\Phi_{i,j} : \mathcal{D}(X_i) \times \mathcal{D}(X_j) \to \mathbb{R}^+ \cup \{0\}$ ($\mathcal{D}(X)$ domain of X)
- $\mathcal{P}(X_1, \ldots X_n) = \frac{1}{Z} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$

- $Z = \sum_{X_1,\ldots X_n} \prod_{i,j} \Phi_{i,j}(X_i, X_j)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

Pairwise MRF: an example



$\mathbf{\Phi_{A,B}}$		$B=0$	$B=1$
A	0	10	1
	1	1	10

$\mathbf{\Phi_{B,C}}$		$C=0$	$C=1$
B	0	10	1
	1	1	10

$\mathbf{\Phi_{A,C}}$		$C=0$	$C=1$
A	0	10	1
	1	1	10

(A,B,C)	$\Phi_{A,B}$	$\Phi_{B,C}$	$\Phi_{A,C}$	$\hat{\mathcal{P}}(A,B,C)$	$\frac{1}{Z}\hat{\mathcal{P}}(A,B,C)$
$(0,0,0)$	10	10	10	1000	0.485
$(0,0,1)$	10	1	1	10	0.005
$(0,1,0)$	1	1	10	10	0.005
$(0,1,1)$	1	10	1	10	0.005
$(1,0,0)$	1	10	1	10	0.005
$(1,0,1)$	1	1	10	10	0.005
$(1,1,0)$	10	1	1	10	0.005
$(1,1,1)$	10	10	10	1000	0.485
			$\sum$	Z=2060	1

$$Z = \sum_{a,b,c\in\{0,1\}^3} \Phi_{A,B}(a,b)\Phi_{B,C}(bc)\Phi_{A,C}(a,c)$$

Question

Is a pairwise MRF able to model any probability distribution over n variables?

- A possibility could be that of buiding a complete graph over the n variables
- The number of required parameters for the joint distribution of n variables with d states and no independence assumption is $O(n^d)$
- However: number of arcs is $\binom{n}{2}$; number of states of the variables is d, then the required number of parameters is $O(n^2 d^2)$
- Pairwise MRF cannot specify the required number of parameters ($n^d \gg n^2 d^2$)

Markov Random Field: Gibbs Distribution

A Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each maximal cliques $C_i \subseteq V$ is associated with a factor or potential $\Phi_i : \mathcal{D}(X_1^i) \times \ldots \times \mathcal{D}(X_k^i) \to \mathbb{R}^+ \cup \{0\}$ where $X_1^i \ldots X_k^i$ are the vertices in C_i;
- $\mathcal{P}(X_1, \ldots X_n) = \frac{1}{Z} \prod_{C_i} \Phi_i(X_1^i, \ldots X_k^i)$

- $Z = \sum_{X_1,\ldots X_n} \prod_i \Phi_i(X_1^i, \ldots X_k^i)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

Markov Random Field: Gibbs Distribution

A Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each maximal cliques $C_i \subseteq V$ is associated with a factor or potential $\Phi_i : \mathcal{D}(X_1^i) \times \ldots \times \mathcal{D}(X_k^i) \to \mathbb{R}^+ \cup \{0\}$ where $X_1^i \ldots X_k^i$ are the vertices in C_i;
- $\mathcal{P}(X_1, \ldots X_n) = \frac{1}{Z} \prod_{C_i} \Phi_i(X_1^i, \ldots X_k^i)$

- $Z = \sum_{X_1,\ldots X_n} \prod_i \Phi_i(X_1^i, \ldots X_k^i)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

Markov Random Field: Gibbs Distribution

A Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \dots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each maximal cliques $C_i \subseteq V$ is associated with a factor or potential $\Phi_i : \mathcal{D}(X_1^i) \times \dots \times \mathcal{D}(X_k^i) \to \mathbb{R}^+ \cup \{0\}$ where $X_1^i \dots X_k^i$ are the vertices in C_i;
- $\mathcal{P}(X_1, \dots X_n) = \frac{1}{Z} \prod_{C_i} \Phi_i(X_1^i, \dots X_k^i)$

- $Z = \sum_{X_1,\dots X_n} \prod_i \Phi_i(X_1^i, \dots X_k^i)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

Markov Random Field: Gibbs Distribution

A Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \dots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each maximal cliques $C_i \subseteq V$ is associated with a factor or potential $\Phi_i : \mathcal{D}(X_1^i) \times \dots \times \mathcal{D}(X_k^i) \to \mathbb{R}^+ \cup \{0\}$ where $X_1^i \dots X_k^i$ are the vertices in C_i;
- $\mathcal{P}(X_1, \dots X_n) = \frac{1}{Z} \prod_{C_i} \Phi_i(X_1^i, \dots X_k^i)$

- $Z = \sum_{X_1, \dots X_n} \prod_i \Phi_i(X_1^i, \dots X_k^i)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

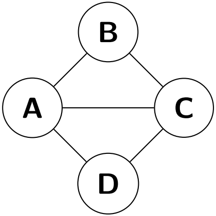
Markov Random Field: Gibbs Distribution

A Markov Random Field is a pair $\langle G, \mathcal{P} \rangle$ where:

- $G = \langle V, E \rangle$ is an undirected graph whose vertices $V = \{X_1, \ldots X_n\}$ represent (discrete) random variables and and edge $(X_i - X_j) \in E$ represents a dependence between X_i and X_j
- each maximal cliques $C_i \subseteq V$ is associated with a factor or potential $\Phi_i : \mathcal{D}(X^i_1) \times \ldots \times \mathcal{D}(X^i_k) \rightarrow \mathbb{R}^+ \cup \{0\}$ where $X^i_1 \ldots X^i_k$ are the vertices in C_i;
- $\mathcal{P}(X_1, \ldots X_n) = \frac{1}{Z} \prod_{C_i} \Phi_i(X^i_1, \ldots X^i_k)$

- $Z = \sum_{X_1, \ldots X_n} \prod_i \Phi_i(X^i_1, \ldots X^i_k)$ and is called the partition function (normalization factor)
- we say that $\mathcal{P}$ factorizes over G

Gibbs Distribution MRF: an example



Max Cliques: $C_1 = \{A, B, C\}$; $C_2 = \{A, C, D\}$

(A, B, C)	$\Phi_1(A, B, C)$
$(0, 0, 0)$	100
$(0, 0, 1)$	1
$(0, 1, 0)$	100
$(0, 1, 1)$	1
$(1, 0, 0)$	1
$(1, 0, 1)$	100
$(1, 1, 0)$	1
$(1, 1, 1)$	100

(A, C, D)	$\Phi_2(A, C, D)$
$(0, 0, 0)$	100
$(0, 0, 1)$	1
$(0, 1, 0)$	100
$(0, 1, 1)$	1
$(1, 0, 0)$	1
$(1, 0, 1)$	100
$(1, 1, 0)$	1
$(1, 1, 1)$	100

Probabilistic Graphical Models
Luigi Portinale
Probabilistic Graphical Models
Bayesian Networks
Markov Random Fields
Conditional Independence and Factorization
Bayesian Networks
Markov Random Fields
BN and MRF: Common Notions on Independence
Conversion between BN and MRF
Factor Graphs

Gibbs Distribution

(A,B,C,D)	Φ_1	Φ_2	$\hat{\mathcal{P}}(A,B,C)$	$\frac{1}{Z}\hat{\mathcal{P}}(A,B,C)$
$(0,0,0,0)$	100	100	10000	0.245
$(0,0,0,1)$	100	1	100	0.002
$(0,0,1,0)$	1	100	100	0.002
$(0,0,1,1)$	1	1	1	2.4×10^{-05}
$(0,1,0,0)$	100	100	10000	0.245
$(0,1,0,1)$	100	1	100	0.002
$(0,1,1,0)$	1	100	100	0.002
$(0,1,1,1)$	1	1	1	2.4×10^{-05}
$(1,0,0,0)$	1	1	1	2.4×10^{-05}
$(1,0,0,1)$	1	100	100	0.002
$(1,0,1,0)$	100	1	100	0.002
$(1,0,1,1)$	100	100	10000	0.245
$(1,1,0,0)$	1	1	1	2.4×10^{-05}
$(1,1,0,1)$	1	100	100	0.002
$(1,1,1,0)$	100	1	100	0.002
$(1,1,1,1)$	100	100	10000	0.245
		$\sum$	**Z=40804**	**1**

$$Z = \sum_{a,b,c,d \in \{0,1\}^4} \Phi_1(a,b,c)\Phi_2(a,c,d)$$

Conditional Independence and Factorization

Bayesian Networks

v-structure

Given a DAG $G = \langle V, E \rangle$ and un undirected path (*trail*) $X_1 \dots X_n$, there is a v-structure in the trail if there exists a node X_i in the trail such that X_{i-1} and X_{i+1} are parents of X_i in G (i.e., $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$)

Active trail

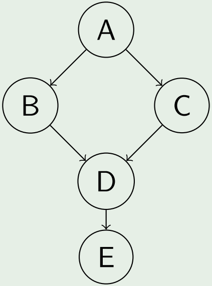
Given a DAG $G = \langle V, E \rangle$, and a subset of nodes $Z \subset V$, a trail $X_1 \dots X_n$ is said to be active given Z if

- for any v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, we have that X_i or one of its descendant is in Z;
- no other X_i along the trail and not in a v-structure is in Z

d-Separation

Given a DAG $G = \langle V, E \rangle$, two nodes $X, Y \in V$ are said to be d-separated given $Z \subset V$ if there is no active trail between X and Y given Z. We indicate this through the notation $dsep_G(X, Y|Z)$

Example



$dsep_G(A, D|B, C)$
$dsep_G(B, C|A)$
$\neg dsep_G(B, C|A, D)$
$\neg dsep_G(B, C|A, E)$
$desp_G(A, E|D)$

I-map

Given a DAG $G = \langle V, E \rangle$ and a probability distribution $\mathcal{P}$ over the variables/nodes of G, we say that G is an I-map (Independence Map) of $\mathcal{P}$ if and only if it satisfies $\mathcal{I}(G) = \{(X \perp Y | Z) : dsep_G(X, Y|Z)\}$

- i.e., d-separation captures actual conditional independencies (there may be more independencies in $\mathcal{P}$ not captured by d-separation)
- we call $\mathcal{I}(G)$ the global independence property

Factorization in BN

Let G be a DAG over variables $X_1 \dots X_n$, a distribution $\mathcal{P}$ over the same space of variables factorizes according to G if it can be expressed as follows

$$\mathcal{P}(X_1 \dots X_n) = \prod_{i=1}^{n} \mathcal{P}(X_i | \pi(X_i))$$

with $\pi(X) = \{Y \in V : Y \text{ is a parent of } X \text{ in } G\}$

Theorem: Independence and Factorization

A distribution $\mathcal{P}$ factorizes according to a DAG G if and only if G is an I-map for $\mathcal{P}$.

Part of previous theorem states the soundness of d-separation. Let $\mathcal{I}(\mathcal{P})$ be the set of independencies present in distribution $\mathcal{P}$

Soundness of d-separation

If $\mathcal{P}$ factorizes according to G then G is an I-map of $\mathcal{P}$; i.e., $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{P})$.

Question: is d-separation detecting ALL the independencies of $\mathcal{I}(\mathcal{P})$? i.e., is it also true that $\mathcal{I}(\mathcal{P}) \subseteq \mathcal{I}(\mathcal{G})$?

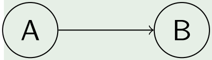
Unfortunately it turns out that $\mathcal{I}(\mathcal{G}) \neq \mathcal{I}(\mathcal{P})$, meaning that even if $\mathcal{P}$ factorizes according to G, it may contain dependencies that cannot be captured by the structure of G.

Faithfulness

A distribution $\mathcal{P}$ is faithful to G if whenever $(X \perp Y | Z) \in \mathcal{I}(\mathcal{P})$, then $dsep_G(X, Y|Z)$.

If any distribution that factorizes over G would be faithful to G, then we would prove the *completeness* of d-separation; i.e., if $\neg dsep_G(X, Y|Z)$ then X and Y would be dependent given Z in all distribution that factorize over G.

Counterexample



$$P(B = b_0|A = a_0) = P(B = b_0|A = a_1)$$
$$P(B = b_1|A = a_0) = P(B = b_1|A = a_1)$$
$(A \perp B)$ but $\neg dsep_G(A, B|\emptyset)$

Weak Completeness

Let G be a DAG; if $\neg dsep_G(X, Y|Z)$ then X and Y are dependent given Z in some distribution $\mathcal{P}$ that factorizes according to G.

- the property does not hold for any possible distribution that factorizes over G
- alternative forumulation: if $(X \perp Y|Z)$ for all the distributions $\mathcal{P}$ that factorize according to G, then $dsep_G(X, Y|Z)$.

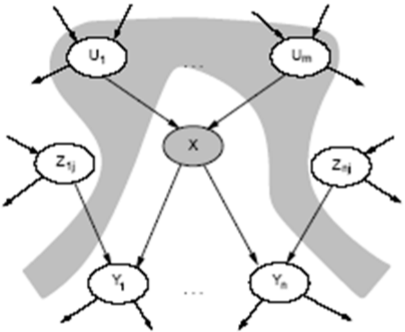
Remark

For almost all the distributions $\mathcal{P}$ that factorizes over G, we have that $\mathcal{I}(\mathcal{P}) = \mathcal{I}(G)$

BN: Local Independence Criteria

Local Independence

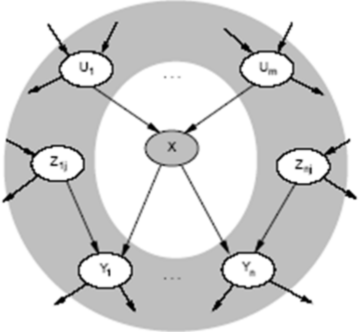
Given a BN $N = \langle G, \mathcal{P} \rangle$ we define the local independence property as $\mathcal{I}_l(N) = \{(X \perp ND_X | \pi(X)\}$ where ND_X are the non-descendant nodes of X in G and $\pi(X)$ are the parents of X in G. *Any node X is independent from its non-descendant given its parents.*



Markov Blanket

Given a BN $N = \langle G, \mathcal{P} \rangle$ and a node X, the Markov Blanket of X, denoted as $MB(X)$ is the set of *parents, children* and *mates* (other parents of the children) of X in G.

Any node X is independent from the rest of the network, given $MB(X)$.



Local independence property and Markov Blanket independence are equivalent to the global semantics i.e., that $\mathcal{P}$ factorizes over G

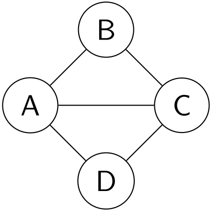
Markov Random Fields

Active path

Given an MRF with graph $G = \langle V, E \rangle$, we say that a path $X_1 - \ldots - X_k$ in G is active given $Z \subset V$, if no $X_i (1 \leq i \leq k)$ is in Z

Separation

Given an MRF with graph $G = \langle V, E \rangle$, two nodes $X, Y \in V$ and $Z \subset V$, we say that X is separated from Y given Z (or alternatively that Z separates X from Y), denoted $sep_G(X, Y|Z)$, if there is no active path connecting X and Y given Z



$sep_G(B, D|AC)$ $\neg sep_G(B, D|A)$ $\neg sep_G(B, D|C)$

I-map

Given a graph G and a probability distribution $\mathcal{P}$ over the variables/nodes of G, we say that G is an I-map (Independence Map) of $\mathcal{P}$ if and only if $\mathcal{P}$ satisfies
$\mathcal{I}(G) = \{(X \perp Y | Z) : sep_G(X, Y | Z)\}$

- i.e., separation captures actual conditional independencies (there may be however independencies in $\mathcal{P}$ not captured by separation)
- we call $\mathcal{I}(G)$ the global independence property

Factorization in MRF

Given a distribution $\mathcal{P}_\Phi$ with factors $\Phi = \{\Phi_1(D_1), \ldots \Phi_k(D_k)\}$ (i.e. $\mathcal{P}_\Phi = \prod_{i=1}^{k} \Phi_i(D_i)$) and an MRF $\langle G, \mathcal{P} \rangle$, we say that $\mathcal{P}_\Phi$ factorizes on G if and only if each D_i $(1 \leq i \leq k)$ is a clique (complete subgraph) of G; in other words $\mathcal{P} = \mathcal{P}_\Phi$.

- each factor Φ_i is calles a clique potential
- without lost of generality, we consider cliques as maximal cliques (i.e. complete subgraph that cannot be extended to a clique by adding adjacent nodes)

A product of factors

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

$\Rightarrow$

a^1	b^1	c^1	$0.5 \cdot 0.5 = 0.25$
a^1	b^1	c^2	$0.5 \cdot 0.7 = 0.35$
a^1	b^2	c^1	$0.8 \cdot 0.1 = 0.08$
a^1	b^2	c^2	$0.8 \cdot 0.2 = 0.16$
a^2	b^1	c^1	$0.1 \cdot 0.5 = 0.05$
a^2	b^1	c^2	$0.1 \cdot 0.7 = 0.07$
a^2	b^2	c^1	$0 \cdot 0.1 = 0$
a^2	b^2	c^2	$0 \cdot 0.2 = 0$
a^3	b^1	c^1	$0.3 \cdot 0.5 = 0.15$
a^3	b^1	c^2	$0.3 \cdot 0.7 = 0.21$
a^3	b^2	c^1	$0.9 \cdot 0.1 = 0.09$
a^3	b^2	c^2	$0.9 \cdot 0.2 = 0.18$

Soundness

Let $\mathcal{P}$ be a distribution over V and G the graph of an MRF with nodes V; if $\mathcal{P}$ factorizes over G, then G is an I-map for $\mathcal{P}$.

- the separation criterion is sound wrt conditional independencies in $\mathcal{P}$

Hammersley-Clifford Theorem

If $\mathcal{P}$ is a positive distribution over V and G is the graph of an MRF with nodes V, if G is an I-map for $\mathcal{P}$, then $\mathcal{P}$ factorizes over G.

- conditional independencies represented in the graph allows the factorization of the distribution (but only if it is positive!)

a positive distribution $\mathcal{P}$ factorizes over a graph G if and only if G is an I-map for $\mathcal{P}$

Weak Completeness

If X and Y are not separated given Z in a graph G of an MRF, then X and Y are dependent given Z in some distribution $\mathcal{P}$ that factorizes over G.

- the above property does not hold in general for any possibile distribution that factorizes over G

Remark

As for BN, for almost all the distributions $\mathcal{P}$ that factorizes over G, we have that $\mathcal{I}(\mathcal{P}) = \mathcal{I}(G)$

Probabilistic Graphical Models
Luigi Portinale
Probabilistic Graphical Models
Bayesian Networks
Markov Random Fields
Conditional Independence and Factorization
Bayesian Networks
Markov Random Fields
BN and MRF: Common Notions on Independence
Conversion between BN and MRF
Factor Graphs

MRF: Local Independence Criteria

Pairwise independence

Given an MRF with graph $G = \langle V, E \rangle$, we define the pairwise independency property as
$\mathcal{I}_p(G) = \{(X \perp Y | V - \{X, Y\}) : (X - Y \notin E)\}$

- Any two non-adjacent variables are conditionally independent given all other variables.

Markov Blanket Independence

The Markov Blanket $MB_G(X)$ of a node X in a graph G is the set of all neighbors of X in G; we define the local independency property as
$\mathcal{I}_l(G) = \{(X \perp V - \{X\} - MB_G(X) | MB_G(X)) : X \in V\}$

- A node is conditionally independent of all the rest of the nodes given its immediate neighbors.

Relationships between independence properties

- pairwise independence is strictly weaker than local independence that is strictly weaker than global independence
  - if $\mathcal{P} \models \mathcal{I}_l(G)$ then $\mathcal{P} \models \mathcal{I}_p(G)$
  - if $\mathcal{P} \models \mathcal{I}(G)$ then $\mathcal{P} \models \mathcal{I}_l(G)$
- in case of positive distribution we also have that if $\mathcal{P} \models \mathcal{I}_p(G)$ then $\mathcal{P} \models \mathcal{I}(G)$

In case of positive distributions all the properties are equivalent

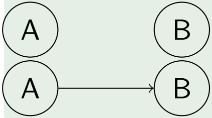
BN and MRF: Common Notions on Independence

- A generic I-map cannot be so useful e.g., a complete graph (either undirected or a DAG) does not imply any independence, thus is an I-map for any distribution
- Recall: an I-map is such that is there is a (d)separation, then it corresponds to a conditional independence; if there is no such a (d)separation (as in a complete graph) than the definition is satisfied.

Minimal I-map

An I-map is minimal if the removal of even a single edge renders it not an I-map

Example



$$P(B|A = a_0) = P(B|A = a_1)$$

Both are I-maps, but the second is not minimal.

BN: Minimal I-map Construction

- Fix an ordering $X_1 \dots X_n$;
- for each i
  - select $\pi(X_i)$ the minimal set $\{X_1 \dots X_{i-1}\}$ such that $(X_1 \perp \{X_i \dots X_{i-1}\} - \pi(X_i) | \pi(X_i))$

MRF: Minimal I-map Construction

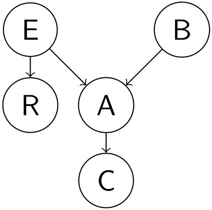
For positive distributions, the simplest way is to exploit pairwise independence.

- Given a a distribution $\mathcal{P}$ over a set of variable $\mathcal{X}$, add an edge between any pair $X, Y \in \mathcal{X}$ such that $\mathcal{P} \not\models (X \perp Y | \mathcal{X} - \{X, Y\})$ (i.e., $\neg(X \perp Y | \mathcal{X} - \{X, Y\})$)

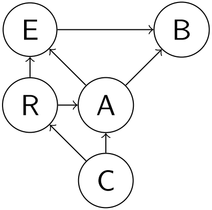
Similar construction can be defined for MB independence

BN: Uniqueness of minimal I-map

In a BN there is no unique minimal I-maps (different orderings of the variables applied to the construction procedure may produce different minimal I-maps)



Order: E,B,A,C,R



Order: C, R, A, E, B

Suppose left graph encodes exactly the probabilitic dependencies of our target distribution (we see later it is a P-map)

$(E \perp B)$ $(E \perp B|R)$ $(E \perp C|A)$ $(B \perp R)$ $(B \perp R|A, E)$ $(B \perp R|C, E)$ $(B \perp R|A, C, E)$ $(B \perp C|A)$ $(A \perp R|E)$ $(C \perp R|E)$ etc...

MRF: Uniqueness of minimal I-map

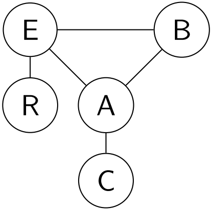
Let $\mathcal{P}$ be a positive distribution and G be the graph constructed on the variables of $\mathcal{P}$ by adding an edge for all X, Y satisfying the pairwise independence property, then G is the unique minimal I-map of $\mathcal{P}$

The same uniqueness result holds if you use the MB independence property.

Remark

A BN has no unique minimal I-map for its distribution.
An MRF has a unique minimal I-map for its distribution

Minimal I-map for the distribution of E, B, R, A, C



It is easy to see that if we add edges following the pairwise independence property, then we end up with this network (e.g., when considering the pair of nodes E, B, we test if $(E \perp B | R, A, C)$; since it turns out to be false, then we add an edge between E and B)

P-map

A graph G is a P-map (Perfect Map) for a distribution $\mathcal{P}$ if $\mathcal{I}(G) = \mathcal{I}(\mathcal{P})$

Unfortunately P-maps may not exist!

Question n.1

Is there a BN that is a perfect map for a given MRF? (*answer: NO, diamond network*)

Question n.2

Is there an MRF that is a perfect map for a given BN? (*answer: NO, v-structure*)

P-map

A graph G is a P-map (Perfect Map) for a distribution $\mathcal{P}$ if $\mathcal{I}(G) = \mathcal{I}(\mathcal{P})$

Unfortunately P-maps may not exist!

Question n.1

Is there a BN that is a perfect map for a given MRF? (*answer: NO, diamond network*)

Question n.2

Is there an MRF that is a perfect map for a given BN? (*answer: NO, v-structure*)

P-map

A graph G is a P-map (Perfect Map) for a distribution $\mathcal{P}$ if $\mathcal{I}(G) = \mathcal{I}(\mathcal{P})$

Unfortunately P-maps may not exist!

Question n.1

Is there a BN that is a perfect map for a given MRF? (*answer: NO, diamond network*)

Question n.2

Is there an MRF that is a perfect map for a given BN? (*answer: NO, v-structure*)

P-map

A graph G is a P-map (Perfect Map) for a distribution $\mathcal{P}$ if $\mathcal{I}(G) = \mathcal{I}(\mathcal{P})$

Unfortunately P-maps may not exist!

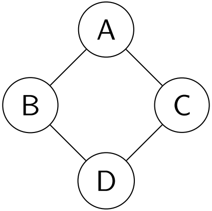
Question n.1

Is there a BN that is a perfect map for a given MRF? (*answer: NO, diamond network*)

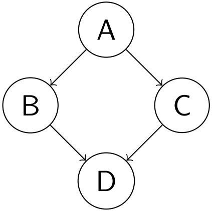
Question n.2

Is there an MRF that is a perfect map for a given BN? (*answer: NO, v-structure*)

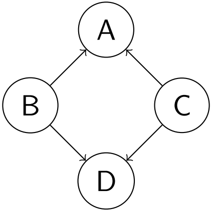
MRF Diamond Network



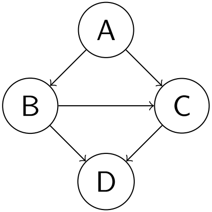
$(A \perp D | B, C)$; $(B \perp C | A, D)$



$(B \perp C | A)$; not an I-map

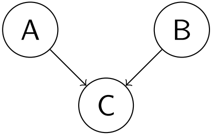


$(B \perp C)$; not an I-map

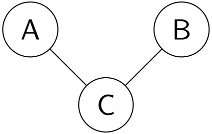


$(A \perp D | B, C)$; $\mathcal{I}(G) \subset \mathcal{I}(\mathcal{P})$

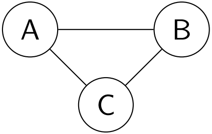
BN v-structure



$(A \perp B); \neg(A \perp B|C)$



$(A \perp B|C)$



$\neg(A \perp B|C); \neg(A \perp B)$

Uniqueness of a P-map

A P-map if it exists, it is not genarally unique. Multiple graphs can encode precisely the same independence assumptions.

- They both encode $\neg(A \perp B)$ and are perfect map for such property.
- however, if multiple P-maps exist, they are I-equivalent

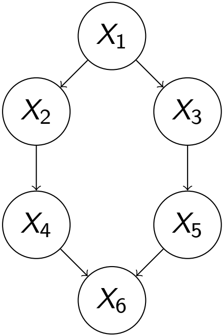
I-equivalence

Two graph G_1 and G_2 are *I-equivalent* iff $\mathcal{I}(G_1) = \mathcal{I}(G_2)$ (i.e., they represent the same independence assumptions)

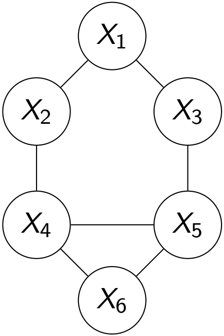
Conversion between BN and MRF

From BN to MRF

- We can build an MRF such that $\mathcal{I}(MRF) \subseteq \mathcal{I}(BN)$
- We cannot just remove arrows at the arcs, since we would add independencies when a v-structure is present
- **Solution**: remove arrows and then moralize (i.e., connect nodes having a common child, meaning every node is then connected to its Markov Blanket). Assign each CPT to one clique potential that contains it.



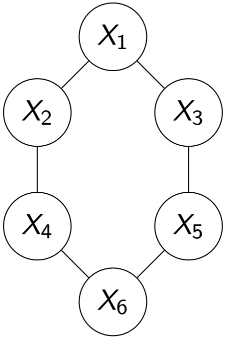
$(X_4 \perp X_5 | X_1)$



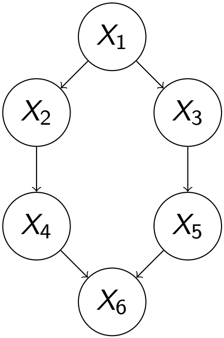
$\neg(X_4 \perp X_5 | X_1)$

From MRF to BN

- We can build a BN such that $\mathcal{I}(BN) \subseteq \mathcal{I}(MRF)$
- We cannot just add arrows at the arcs, since we would add independencies when a v-structure is present



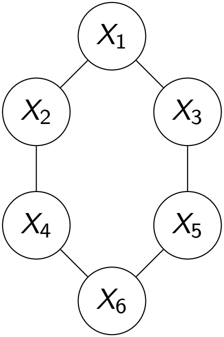
$\neg(X_2 \perp X_3 | X_1)$



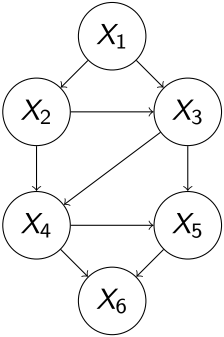
$(X_2 \perp X_3 | X_1)$

From MRF to BN (cont.)

- **Solution**: Fix an order of the nodes, then add each node along with its minimal parent set according to the independencies defined in the MRF



$(X_1 \perp X_6 | X_2, X_5)$

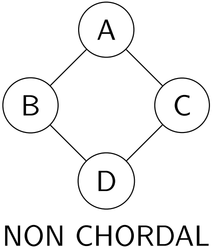


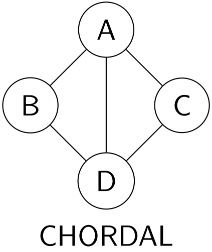
$\neg(X_1 \perp X_6 | X_2, X_5)$

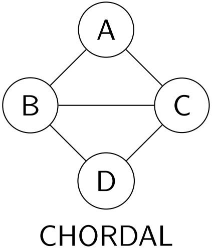
When can BN and MRF model the same distribution?

Chordal Graph

An undirected graph G is chordal or triangulated iff any loop of length greater than 3 has a chord. A chord in a loop is an edge connecting two non-consecutive nodes in the loop.
A directed graph is said to be chordal iff the underlying undirected graph is chordal.



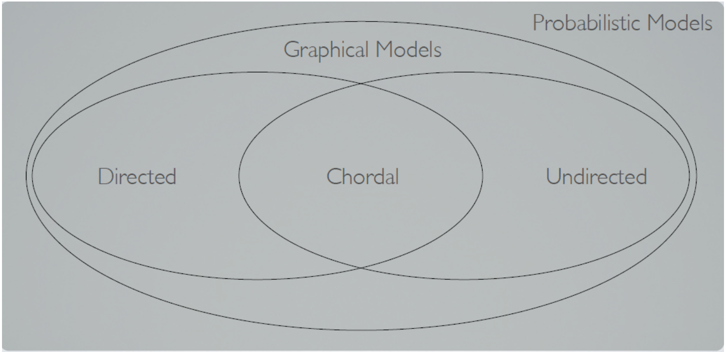




Theorem

Let G be the graph of a chordal MRF, then it exist a BN whose graph H is such that $\mathcal{I}(G) = \mathcal{I}(G)$

Type of PGM



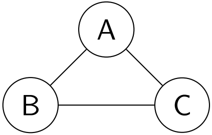
Factor Graphs

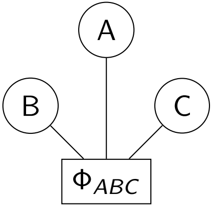
- Factor graphs are a representation that makes explicit how the factorization takes place on a graph (relevant for MRF).

Factor Graph

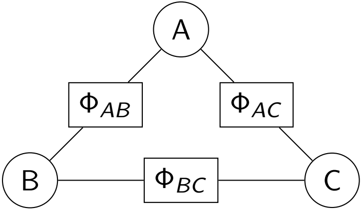
A factor graph (FG) F is an undirected bi-partite graph where nodes are of type *variables* (circles) and *factors* (squares); an edge is present between a variable and a factor node iff the variable is contained in the scope of the factor. A distribution $\mathcal{P}$ factorizes over F if it can be represented as a set of factors in this form.

Example





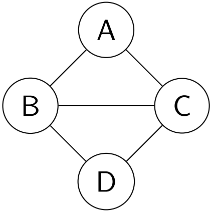
Max cliques factorization



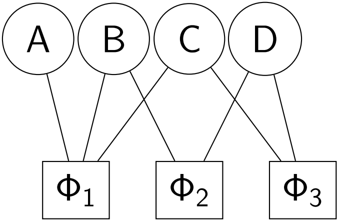
Pairwise factorization

- Remember: given a set of factors we can induce an MRF by just adding an edge between variables which are in the scope of the same factor
- However, given an MRF different factorizations are possible (see previous slide)
- We CANNOT read the factorization from the graph
- Factor Graphs make explicit the factorization

Example: $\Phi_1(A, B, C), \Phi_2(B, C, D)$



However also the following can be factorizations from such a graph: $\Phi_1(A, B, C), \Phi_2(B, D), \Phi_3(C, D)$;
$\Phi_1(A, B), \Phi_2(A, C), \Phi_3(B, C), \Phi_4(B, D), \Phi_5(C, D)$, etc ...



$$\Phi_1(A, B, C), \Phi_2(B, D), \Phi_3(C, D)$$