Probabilistic Graphical Models

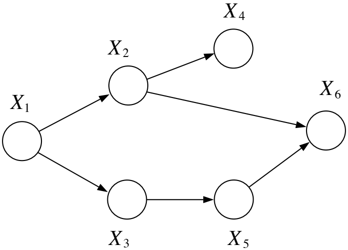
Lecture 3: Undirected Graphical Models

Matthew Walter

TTI-Chicago

April 14, 2020

Bayesian Networks (Revisited)



- G encodes the following **local** independencies

$$I_l(G) = (X_i \perp \mathrm{NonDescendants}_{X_i} \,|\, \mathrm{Pa}^G_{X_i}) \quad \forall X_i$$

- G encodes the following **global (Markov)** independencies

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

Bayesian Networks (Revisited)

- A distribution P **factorizes** over 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), one per node $P(X_i \mid \mathrm{Pa}^G_{X_i})$ specifying probability conditioned on parents
- A graph G encodes conditional independencies that hold under the corresponding distribution P

Representation Theorem (Revisited)

- Let $I(P)$ be the set of independence assertions $\{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})\}$ that hold in P ($P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$)
- 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
  - P may have additional independencies not reflected in G
  - Trivial case: A fully connected graph G is an I-map for *any* distribution since $I(G) = \emptyset \subseteq I(P)\ \forall P$

Theorem

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*

Bayesian Networks: Distributions to Graphs

- 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 $I(G) = I(P)$
- Perfect maps are unique up to I-equivalence

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 $X_j \to X_i$ to $\mathcal{G}$
  return $\mathcal{G}$

Bayesian Networks: Summary

A **Bayesian Network** (G, P) consists of a DAG G the corresponding distribution P, where

- **Factorization**: P that factorizes G as a product of conditional probability distributions (CPDs)
- **Conditional Independence Semantics**: Local and global independencies
- **Conditional Independence Queries**: D-separation

- Directed edges express causality relationships
- Exploiting independencies is essential to inference
- A BN can be viewed as a **generative model**, where variables are sampled in topological order

Misconception Example

- Suppose that there is a secret going around among a group of four people, modeled with the following boolean random variables:
  - A: Andrew knows the secret
  - B: Bob knows the secret
  - C: Carrie knows the secret
  - D: Dennis knows the secret
- Andrew and Dennis are each friends with Bob and Carrie
- Andrew and Dennis don't get along
- Friends only share the secret with one another

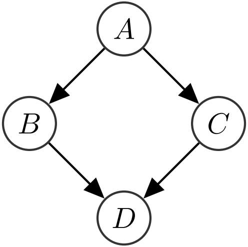
How do we represent these independencies?

$$\{(A \perp D \mid B, C), (B \perp C \mid A, D)\}$$

Misconception Example

We want a graph that captures the following independencies

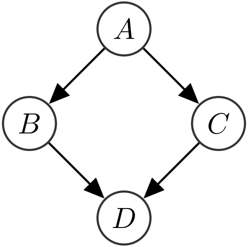
$$\{(A \perp D \mid B, C), (B \perp C \mid A, D)\}$$



Misconception Example

We want a graph that captures the following independencies

$$\{(A \perp D \mid B, C), (B \perp C \mid A, D)\}$$

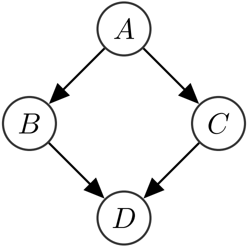


$$(A \perp D \mid B, C)$$
$$(B \not\perp C \mid A, D)$$
$$(B \perp C \mid A)$$

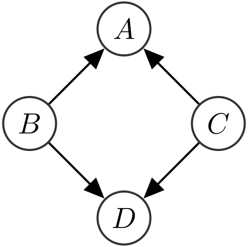
Misconception Example

We want a graph that captures the following independencies

$$\{(A \perp D \mid B, C), (B \perp C \mid A, D)\}$$



$(A \perp D \mid B, C)$
$(B \not\perp C \mid A, D)$
$(B \perp C \mid A)$

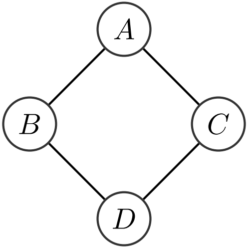


$(A \perp D \mid B, C)$
$(B \not\perp C \mid A, D)$
$(B \perp C)$

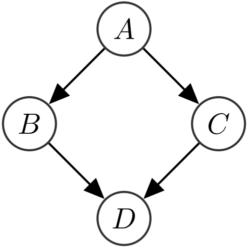
Misconception Example

We want a graph that captures the following independencies

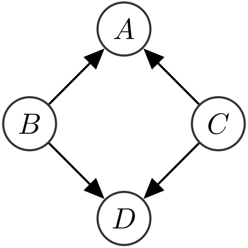
$$\{(A \perp D \mid B, C), (B \perp C \mid A, D)\}$$



$(A \perp D \mid B, C)$
$(B \perp C \mid A, D)$



$(A \perp D \mid B, C)$
$(B \not\perp C \mid A, D)$
$(B \perp C \mid A)$



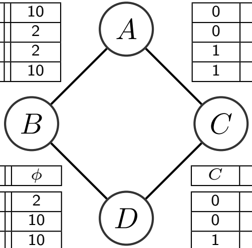
$(A \perp D \mid B, C)$
$(B \not\perp C \mid A, D)$
$(B \perp C)$

Misconception Example: Alternative Formulation

- Suppose that we instead model the relationship between the random variables in terms of their consistency (alt., $\exp(-\text{potential energy})$)

A	B	ϕ
0	0	10
0	1	2
1	0	2
1	1	10

A	C	ϕ
0	0	2
0	1	10
1	0	10
1	1	2



B	D	ϕ
0	0	2
0	1	10
1	0	10
1	1	2

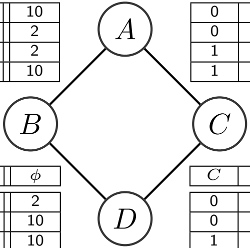
C	D	ϕ
0	0	10
0	1	2
1	0	2
1	1	10

Misconception Example: Alternative Formulation

- Suppose that we instead model the relationship between the random variables in terms of their consistency (alt., $\exp(-\text{potential energy})$)

A	B	ϕ
0	0	10
0	1	2
1	0	2
1	1	10

A	C	ϕ
0	0	2
0	1	10
1	0	10
1	1	2



B	D	ϕ
0	0	2
0	1	10
1	0	10
1	1	2

C	D	ϕ
0	0	10
0	1	2
1	0	2
1	1	10

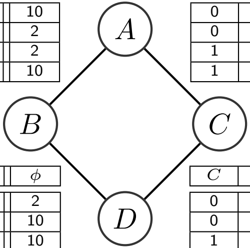
- The overall consistency of a particular setting of the random variables is equivalent to the product of the individual (pair-wise) potentials

Misconception Example: Alternative Formulation

- Suppose that we instead model the relationship between the random variables in terms of their consistency (alt., $\exp(-\text{potential energy})$)

A	B	ϕ
0	0	10
0	1	2
1	0	2
1	1	10

A	C	ϕ
0	0	2
0	1	10
1	0	10
1	1	2

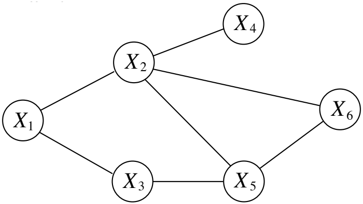


B	D	ϕ
0	0	2
0	1	10
1	0	10
1	1	2

C	D	ϕ
0	0	10
0	1	2
1	0	2
1	1	10

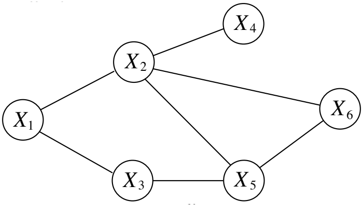
- The overall consistency of a particular setting of the random variables is equivalent to the product of the individual (pair-wise) potentials
- Inference is a problem of maximizing consistency (minimizing energy)

Markov Network Structure



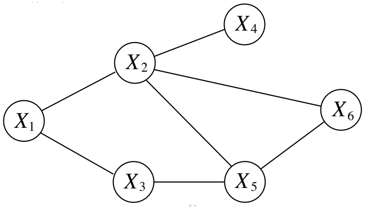
- $H = (V, E)$ is an undirected graph with
  - One node $i \in V$ for each random variable X_i
  - Undirected edges $\{i, j\} \in E$ that capture interactions between random variables X_i and X_j

Markov Network Structure



- $H = (V, E)$ is an undirected graph with
  - One node $i \in V$ for each random variable X_i
  - Undirected edges $\{i, j\} \in E$ that capture interactions between random variables X_i and X_j
- A **clique** $C \subset H$ is a subgraph of H for which there is an edge between every pair of nodes $X_i, X_j \in C$
  - A **maximal clique** is one such that adding any additional node X_k to C will render it no longer a clique

Markov Network Structure



- $H = (V, E)$ is an undirected graph with
  - One node $i \in V$ for each random variable X_i
  - Undirected edges $\{i, j\} \in E$ that capture interactions between random variables X_i and X_j
- A **clique** $C \subset H$ is a subgraph of H for which there is an edge between every pair of nodes $X_i, X_j \in C$
  - A **maximal clique** is one such that adding any additional node X_k to C will render it no longer a clique
- A **factor** is a non-negative potential function over cliques

$$\phi(C) : \mathrm{Val}(C) \rightarrow \mathbb{R}^+$$

Gibbs Distribution

- A distribution P is a **Gibbs distribution** *parameterized* by a set of factors $\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_K(\boldsymbol{D}_K)\}$ over sets $\boldsymbol{D}_k$ if

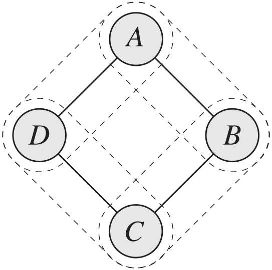
$$P(X_1, \ldots, X_n) = \frac{1}{Z}\phi_1(\boldsymbol{D}_1) \times \ldots \times \phi_K(\boldsymbol{D}_K)$$

where Z is a normalizing constant known as the **partition function**

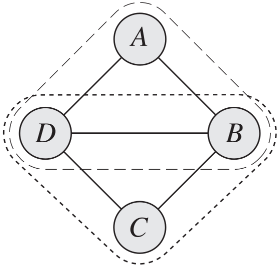
$$Z = \sum_{X_1,\ldots,X_n} \phi_1(\boldsymbol{D}_1) \times \ldots \times \phi_K(\boldsymbol{D}_K)$$

Factorization

- A distribution P with $\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_K(\boldsymbol{D}_K)\}$ **factorizes** over a Markov network H if each $\boldsymbol{D}_k$ is a complete subgraph of H



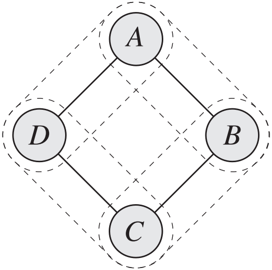
$\{A,B\},\{B,C\},\{C,D\},\{A,D\}$



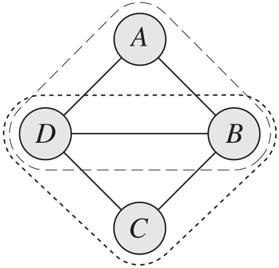
$\{A,B\},\{B,C\},\{C,D\},\{A,D\},\{D,B\}$

Factorization

- A distribution P with $\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_K(\boldsymbol{D}_K)\}$ **factorizes** over a Markov network H if each $\boldsymbol{D}_k$ is a complete subgraph of H
- These factors are often referred to as **clique potentials**



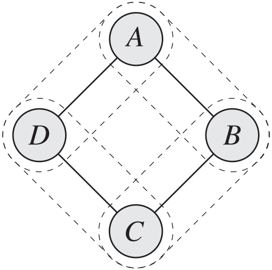
$\{A, B\}, \{B, C\}, \{C, D\}, \{A, D\}$



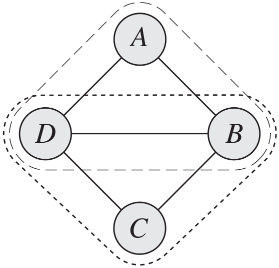
$\{A, B\}, \{B, C\}, \{C, D\}, \{A, D\}, \{D, B\}$

Factorization

- A distribution P with $\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_K(\boldsymbol{D}_K)\}$ **factorizes** over a Markov network H if each $\boldsymbol{D}_k$ is a complete subgraph of H
- These factors are often referred to as **clique potentials**
- We can reduce the number of factors by defining each $\boldsymbol{D}_k$ as a maximal clique (why can we do this? why might we not want to?)



$\{A, B\}, \{B, C\}, \{C, D\}, \{A, D\}$



$\{A, B, D\}, \{B, C, D\}$

Markov Networks (Undirected Graphical Models)

- A **Markov Network** is a pair $B = (P, H)$ for which P is a Gibbs distribution that factorizes over H

$$P(X_i, \ldots, X_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

where Z is the partition function that normalizes the distribution

$$Z = \sum_{X_i, \ldots, X_n} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

Markov Networks (Undirected Graphical Models)

- A **Markov Network** is a pair $B = (P, H)$ for which P is a Gibbs distribution that factorizes over H

$$P(X_i, \ldots, X_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

  where Z is the partition function that normalizes the distribution

$$Z = \sum_{X_i, \ldots, X_n} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

- Provides an alternative representation for joint distributions

Markov Networks (Undirected Graphical Models)

- A **Markov Network** is a pair $B = (P, H)$ for which P is a Gibbs distribution that factorizes over H

$$P(X_i, \ldots, X_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

where Z is the partition function that normalizes the distribution

$$Z = \sum_{X_i, \ldots, X_n} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

- Provides an alternative representation for joint distributions
- Factors are not equivalent to marginal or conditional distributions over variables in their scope

Markov Networks (Undirected Graphical Models)

- A **Markov Network** is a pair $B = (P, H)$ for which P is a Gibbs distribution that factorizes over H

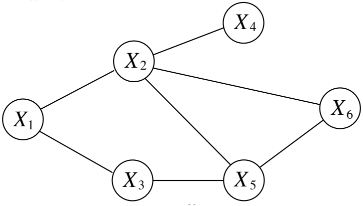
$$P(X_i, \ldots, X_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

  where Z is the partition function that normalizes the distribution

$$Z = \sum_{X_i, \ldots, X_n} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

- Provides an alternative representation for joint distributions
- Factors are not equivalent to marginal or conditional distributions over variables in their scope
- Also known as **Markov random fields** (MRFs)

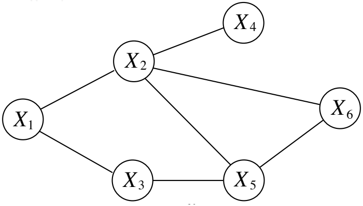
Interpretation of Factors



With Bayesian networks, CPDs provide a local probabilistic interpretation of random variable interactions

Can we also use conditional probabilities for Markov networks?

Interpretation of Factors

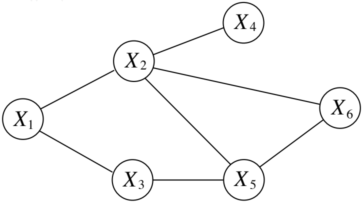


With Bayesian networks, CPDs provide a local probabilistic interpretation of random variable interactions

Can we also use conditional probabilities for Markov networks?

What if we associate with each node the conditional probability of the node given its neighbors?

Interpretation of Factors



With Bayesian networks, CPDs provide a local probabilistic interpretation of random variable interactions

Can we also use conditional probabilities for Markov networks?

What if we associate with each node the conditional probability of the node given its neighbors?

In general, it will be difficult to ensure conditional probabilities are consistent with one another and, thus, the same joint distribution

Interpretation of Factors

$$X \perp Z \,|\, Y$$

- The distribution is $P(X, Y, Z) = P(Y)P(X \,|\, Y)P(Z \,|\, Y)$
- The cliques are $\mathcal{C} = \{\{X, Y\}, \{Y, Z\}\}$
- Possible factors: $\phi_{XY}(X, Y) = P(Y)P(X \,|\, Y)$ & $\phi_{YZ}(Y, Z) = P(Z \,|\, Y)$?
- How about $\phi_{XY}(X, Y) = P(X, Y)$ & $\phi_{YZ}(Y, Z) = P(Y, Z)$?

Interpretation of Factors

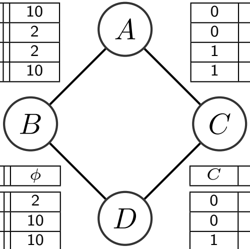
$$X \perp Z \,|\, Y$$

- The distribution is $P(X, Y, Z) = P(Y)P(X \,|\, Y)P(Z \,|\, Y)$
- The cliques are $\mathcal{C} = \{\{X, Y\}, \{Y, Z\}\}$
- Possible factors: $\phi_{XY}(X, Y) = P(Y)P(X \,|\, Y)$ & $\phi_{YZ}(Y, Z) = P(Z \,|\, Y)$?
- How about $\phi_{XY}(X, Y) = P(X, Y)$ & $\phi_{YZ}(Y, Z) = P(Y, Z)$?
- No: $P(X, Y, Z) \neq P(X, Y)P(Y, Z)$ unless $P(Y) = 0$ or $P(Y) = 1$ (i.e., P is not a positive distribution)

Interpretation of Factors

A	B	ϕ
0	0	10
0	1	2
1	0	2
1	1	10

A	C	ϕ
0	0	2
0	1	10
1	0	10
1	1	2



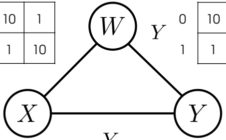
B	D	ϕ
0	0	2
0	1	10
1	0	10
1	1	2

C	D	ϕ
0	0	10
0	1	2
1	0	2
1	1	10

- In general, factors are neither conditional nor marginal probabilities
- Factors (potentials) do not have a probabilistic interpretation
- Instead, they can be interpreted in terms of pre-probabilistic notion of "agreement," "constraint," "consistency," or "energy"

Undirected Graphical Models: Example

Consider a Markov network over binary random variables W, X, and Y



$X \backslash W$	0	1
0	10	1
1	1	10

$Y \backslash W$	0	1
0	10	1
1	1	10

$X \backslash Y$	0	1
0	10	1
1	1	10

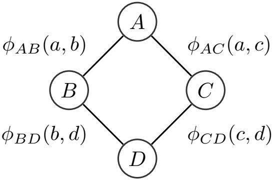
Here, factors encourage equality between each pair of random variables

$$P(W, X, Y) = \frac{1}{Z}\phi_{WX}(W, X) \cdot \phi_{WY}(W, Y) \cdot \phi_{XY}(X, Y)$$

where

$$Z = \sum_{w,x,y \in \{0,1\}^3} \phi_{WX}(w, x) \cdot \phi_{WY}(w, y) \cdot \phi_{XY}(x, Y) = 2 \cdot 1000 + 6 \cdot 10 = 2060$$

Undirected Graphical Models



The joint distribution can be expressed as the product of factors

$$P(a,b,c,d) = \frac{1}{Z}\phi_{AB}(a,b) \cdot \phi_{AC}(a,c) \cdot \phi_{BD}(b,d) \cdot \phi_{CD}(c,d)$$

where

$$Z = \sum_{a,b,c,d \in \{0,1\}^4} \phi_{AB}(a,b) \cdot \phi_{AC}(a,c) \cdot \phi_{BD}(b,d) \cdot \phi_{CD}(c,d)$$

Partition Function

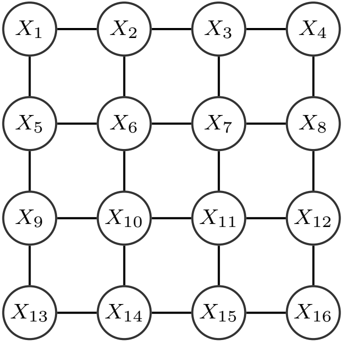
- Term originates (as with Markov random fields) from statistical mechanics
- Can be hard (computationally expensive) to calculate (e.g., summation of 2^n products)

Partition Function

- Term originates (as with Markov random fields) from statistical mechanics
- Can be hard (computationally expensive) to calculate (e.g., summation of 2^n products)
- But ... we don't always need to calculate Z
  - If we are only interested in maximizing $P(X_1, \ldots, X_n)$
  - Related, if we are interested in ratios

$$\frac{P(x_1, \ldots, x_n)}{P(x'_1, \ldots, x'_n)} = \frac{\tilde{P}(x_1, \ldots, x_n)/Z}{\tilde{P}(x'_1, \ldots, x'_n)/Z} = \frac{\tilde{P}(x_1, \ldots, x_n)}{\tilde{P}(x'_1, \ldots, x'_n)}$$

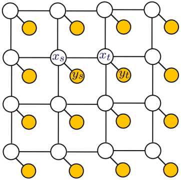
Pairwise Markov Networks



All factors are associated with one or two nodes

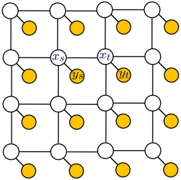
$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i \in V} \phi_i(X_i) \cdot \prod_{\{i,j\} \in E} \phi_{i,j}(X_i, X_j)$$

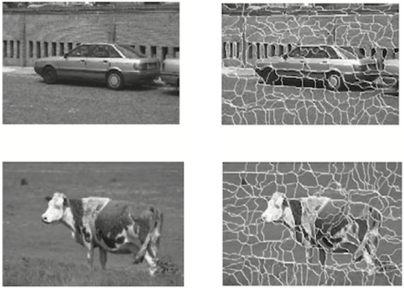
Pairwise Markov Networks: Image Segmentation



Original

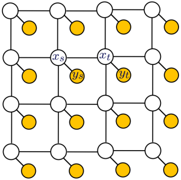
Pairwise Markov Networks: Image Segmentation





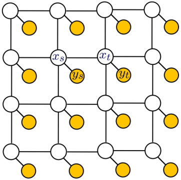
Original	Super pixels

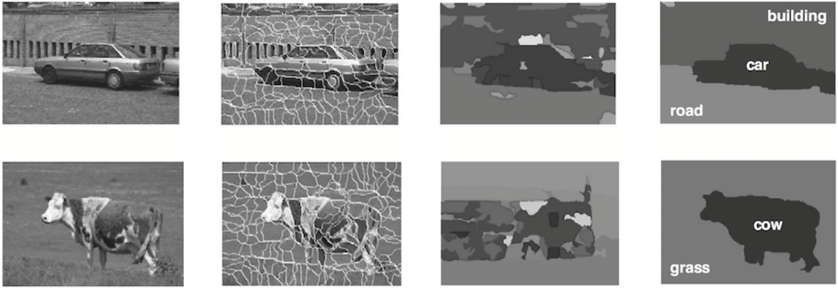
Pairwise Markov Networks: Image Segmentation



Original Super pixels Unary

Pairwise Markov Networks: Image Segmentation





Original | Super pixels | Unary | Unary & Pairwise

Markov Network Independencies

Consider a Markov network structure H over $X_1, \ldots X_n$

- A path $X_i — \cdots — X_k$ in H is **active** given $\boldsymbol{Z}$ if none of the X_i's along the path are in $\boldsymbol{Z}$

Markov Network Independencies

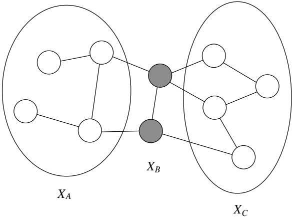
Consider a Markov network structure H over $X_1, \ldots X_n$

- A path $X_i — \cdots — X_k$ in H is **active** given $\boldsymbol{Z}$ if none of the X_i's along the path are in $\boldsymbol{Z}$
- A set of nodes $\boldsymbol{Z}$ **separates** $\boldsymbol{X}$ and $\boldsymbol{Y}$ in H ($\text{sep}_H(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no active path between any $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$

Markov Network Independencies

Consider a Markov network structure H over $X_1, \dots X_n$

- A path $X_i — \cdots — X_k$ in H is **active** given $\boldsymbol{Z}$ if none of the X_i's along the path are in $\boldsymbol{Z}$
- A set of nodes $\boldsymbol{Z}$ **separates** $\boldsymbol{X}$ and $\boldsymbol{Y}$ in H ($\text{sep}_H(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no active path between any $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$

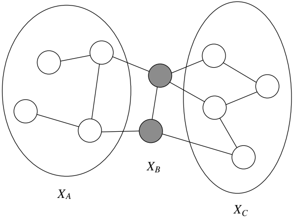


$$\text{sep}_H(X_A; X_C \mid X_B)$$

Markov Network Independencies

Consider a Markov network structure H over $X_1, \ldots X_n$

- A path $X_i — \cdots — X_k$ in H is **active** given $\boldsymbol{Z}$ if none of the X_i's along the path are in $\boldsymbol{Z}$
- A set of nodes $\boldsymbol{Z}$ **separates** $\boldsymbol{X}$ and $\boldsymbol{Y}$ in H ($\text{sep}_H(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no active path between any $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$



$$\text{sep}_H(X_A; X_C \mid X_B)$$

- Separation is monotonic in $\boldsymbol{Z}$, i.e.,

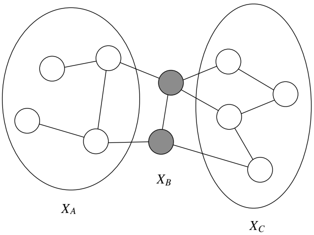
$$\text{for all } \boldsymbol{Z}' \supset \boldsymbol{Z}, \;\; \text{sep}_H(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \Rightarrow \text{sep}_H(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}')$$

Global Markov Independencies

Definition (Global Markov Independencies)

The **Global Markov Independencies** for a Markov network H are

$$I(H) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}) : \mathsf{sep}_H(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})\}$$



Representation Theorem

- **Soundness**: $\mathsf{sep}_H(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) \Rightarrow P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$

Representation Theorem

- **Soundness**: $\text{sep}_H(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \Rightarrow P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$

Theorem

If P is a Gibbs distribution that factorizes over H, then

$$I(H) \subseteq I(P)$$

Representation Theorem

- **Soundness**: $\mathrm{sep}_H(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \Rightarrow P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$

Theorem

If P is a Gibbs distribution that factorizes over H, then

$$I(H) \subseteq I(P)$$

Theorem (Hammersley-Clifford)

If P is a positive distribution and $I(H) \subseteq I(P)$, then P is a Gibbs distribution that factorizes over H

Representation Theorem

- **Soundness**: $\text{sep}_H(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \Rightarrow P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$

Theorem

If P is a Gibbs distribution that factorizes over H, then

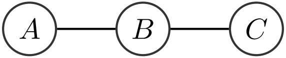
$$I(H) \subseteq I(P)$$

Theorem (Hammersley-Clifford)

If P is a positive distribution and $I(H) \subseteq I(P)$, then P is a Gibbs distribution that factorizes over H

- **Completeness**: $\text{sep}_H(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \Leftarrow P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid Z)$? Not in general

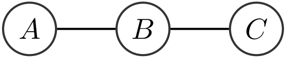
Proof by Example



$$P(a, b, c) = \frac{1}{Z}\phi_{AB}(a, b) \cdot \phi_{BC}(b, c)$$

- We will show that $\text{sep}_H(A; C \mid B) \Rightarrow P \models (A \perp C \mid B)$

Proof by Example



$$P(a, b, c) = \frac{1}{Z}\phi_{AB}(a, b) \cdot \phi_{BC}(b, c)$$

- We will show that $\text{sep}_H(A; C \mid B) \Rightarrow P \models (A \perp C \mid B)$
- First, let's consider $P(a \mid b)$

$$P(a \mid b) = \frac{P(a, b)}{P(b)}$$

Proof by Example

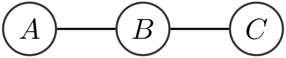
$$\begin{tikzcd} A \arrow[r, no head] & B \arrow[r, no head] & C \end{tikzcd}$$

$$P(a,b,c) = \tfrac{1}{Z}\phi_{AB}(a,b)\cdot\phi_{BC}(b,c)$$

- We will show that $\text{sep}_H(A; C \mid B) \Rightarrow P \models (A \perp C \mid B)$
- First, let's consider $P(a \mid b)$

$$\begin{aligned} P(a \mid b) &= \frac{P(a,b)}{P(b)} \\ &= \frac{\frac{1}{Z}\sum_{c'}\phi_{AB}(a,b)\cdot\phi_{BC}(b,c')}{\frac{1}{Z}\sum_{a',c'}\phi_{AB}(a',b)\cdot\phi_{BC}(b,c')} \end{aligned}$$

Proof by Example

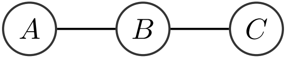


$$P(a,b,c) = \frac{1}{Z}\phi_{AB}(a,b)\cdot\phi_{BC}(b,c)$$

- We will show that $\text{sep}_H(A; C \mid B) \Rightarrow P \models (A \perp C \mid B)$
- First, let's consider $P(a \mid b)$

$$\begin{aligned}
P(a \mid b) &= \frac{P(a,b)}{P(b)} \\
&= \frac{\frac{1}{Z}\sum_{c'}\phi_{AB}(a,b)\cdot\phi_{BC}(b,c')}{\frac{1}{Z}\sum_{a',c'}\phi_{AB}(a',b)\cdot\phi_{BC}(b,c')} \\
&= \frac{\phi_{AB}(a,b)\cdot\sum_{c'}\phi_{BC}(b,c')}{\sum_{a'}\phi_{AB}(a',b)\cdot\sum_{c'}\phi_{BC}(b,c')}
\end{aligned}$$

Proof by Example



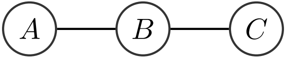
$$P(a,b,c) = \frac{1}{Z}\phi_{AB}(a,b)\cdot\phi_{BC}(b,c)$$

- We will show that $\text{sep}_H(A;C \mid B) \Rightarrow P \models (A \perp C \mid B)$
- First, let's consider $P(a \mid b)$

$$\begin{aligned}
P(a \mid b) &= \frac{P(a,b)}{P(b)} \\
&= \frac{\frac{1}{Z}\sum_{c'}\phi_{AB}(a,b)\cdot\phi_{BC}(b,c')}{\frac{1}{Z}\sum_{a',c'}\phi_{AB}(a',b)\cdot\phi_{BC}(b,c')} \\
&= \frac{\phi_{AB}(a,b)\cdot\sum_{c'}\phi_{BC}(b,c')}{\sum_{a'}\phi_{AB}(a',b)\cdot\sum_{c'}\phi_{BC}(b,c')} = \frac{\phi_{AB}(a,b)}{\sum_{a'}\phi_{AB}(a',b)}
\end{aligned}$$

- More generally, the probability of a variable conditioned on it's neighbors involves only factors over that variable

Proof by Example



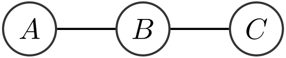
$$P(a, b, c) = \frac{1}{Z}\phi_{AB}(a, b) \cdot \phi_{BC}(b, c)$$

- We will show that $\text{sep}_H(A; C \mid B) \Rightarrow P \models (A \perp C \mid B)$

Proof.

$$P(a, c \mid b) = \frac{p(a, b, c)}{\sum_{a', c'} p(a', b, c')} = \frac{\phi_{AB}(a, b) \cdot \phi_{BC}(b, c)}{\sum_{a', c'} \phi_{AB}(a', b) \cdot \phi_{BC}(b, c')}$$

Proof by Example



$$P(a, b, c) = \frac{1}{Z}\phi_{AB}(a, b) \cdot \phi_{BC}(b, c)$$

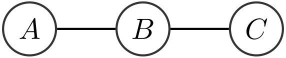
- We will show that $\text{sep}_H(A; C \mid B) \Rightarrow P \models (A \perp C \mid B)$

Proof.

$$\begin{aligned} P(a, c \mid b) = \frac{p(a, b, c)}{\sum_{a', c'} p(a', b, c')} &= \frac{\phi_{AB}(a, b) \cdot \phi_{BC}(b, c)}{\sum_{a', c'} \phi_{AB}(a', b) \cdot \phi_{BC}(b, c')} \\ &= \frac{\phi_{AB}(a, b)}{\sum_{a'} \phi_{AB}(a', b)} \cdot \frac{\phi_{BC}(b, c)}{\sum_{c'} \phi_{BC}(b, c')} \end{aligned}$$

$\square$

Proof by Example



$$P(a,b,c) = \frac{1}{Z}\phi_{AB}(a,b)\cdot\phi_{BC}(b,c)$$

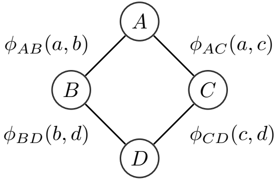
- We will show that $\text{sep}_H(A; C \mid B) \Rightarrow P \models (A \perp C \mid B)$

Proof.

$$\begin{aligned}
P(a,c \mid b) = \frac{p(a,b,c)}{\sum_{a',c'} p(a',b,c')} &= \frac{\phi_{AB}(a,b)\cdot\phi_{BC}(b,c)}{\sum_{a',c'}\phi_{AB}(a',b)\cdot\phi_{BC}(b,c')} \\
&= \frac{\phi_{AB}(a,b)}{\sum_{a'}\phi_{AB}(a',b)} \cdot \frac{\phi_{BC}(b,c)}{\sum_{c'}\phi_{BC}(b,c')} \\
&= P(a \mid b)P(c \mid b)
\end{aligned}$$

$\square$

Communication Example (Revisited)



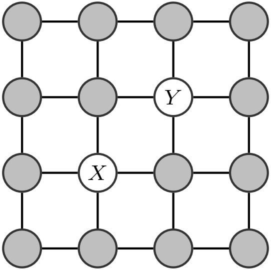
- $\text{sep}_H(A; D \mid B, C) \Rightarrow (A \perp D \mid B, C)$
- $\text{sep}_H(B; C \mid A, D) \Rightarrow (B \perp C \mid A, D)$
- Graph implies no other independencies

Local Markov Independencies

Definition (Pairwise Markov Independencies)

Any pair of non-neighboring variables are independent given everything else

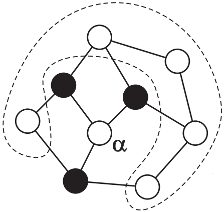
$$I_p(H) = \{(X \perp Y \mid \boldsymbol{X} - \{X, Y\}) : X—Y \notin H\}$$



Local Markov Independencies

Definition (Markov Blanket (Graph))

For an undirected graph H, the **Markov blanket** of a node X $\text{MB}_H(X)$ is the neighbors of X

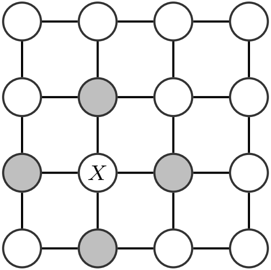


Local Markov Independencies

Definition (Local Markov Independencies)

A random variable X is independent of all other variables given its Markov blanket $\text{MB}_H(X)$

$$I_l(H) = \{(X \perp (\boldsymbol{X} - X - \text{MB}_H(X)) \mid \text{MB}_H(X)) \; : X \in \boldsymbol{X}\}$$



Relationship Between Markov Independencies

- For any Markov network H and distribution P,

$$P \models I_l(H) \Rightarrow P \models I_p(H)$$

since separation is monotonic

Relationship Between Markov Independencies

- For any Markov network H and distribution P,

$$P \models I_l(H) \Rightarrow P \models I_p(H)$$

  since separation is monotonic
- For any Markov network H and distribution P,

$$P \models I(H) \Rightarrow P \models I_l(H)$$

  since separation is monotonic

Relationship Between Markov Independencies

- For any Markov network H and distribution P,

$$P \models I_l(H) \Rightarrow P \models I_p(H)$$

  since separation is monotonic
- For any Markov network H and distribution P,

$$P \models I(H) \Rightarrow P \models I_l(H)$$

  since separation is monotonic
- For a *positive* distribution P,

$$P \models I_p(H) \Rightarrow P \models I(H)$$

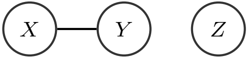
Positive Distribution Requirement: Example

To understand why the distribution is required to be positive, consider an example with binary random variables X, Y, Z

Let P be a distribution such that

$$P(X = 0) = 1/2$$
$$P(X = Y = Z) = 1$$

Let H be:



- $P \models I_p(H) = \{(X \perp Z \mid Y), (Y \perp Z \mid X)\}$
- But, $P \not\models I_l(H)$, which includes $Z \perp X, Y$

Relationship Between Markov Independencies

- For a *positive* distribution P, the following are equivalent
  1. $P \models I_l(H)$
  2. $P \models I_p(H)$
  3. $P \models I(H)$
- Non-positive distributions P (i.e., one or more events has 0 probability) satisfy some (weaker) but not all (stronger) properties

Relationship Between Markov Independencies

- For a *positive* distribution P, the following are equivalent
  1. $P \models I_l(H)$
  2. $P \models I_p(H)$
  3. $P \models I(H)$
- Non-positive distributions P (i.e., one or more events has 0 probability) satisfy some (weaker) but not all (stronger) properties
- This equivalence is useful in constructing Markov networks that are minimal I-maps for a given (positive) distribution

Distributions to Graphs: Pairwise Independencies

- Given distribution P, we would like to construct the minimal I-map H
- Construct H by adding an edge between all pairs X and Y such that

$$P \not\models (X \perp Y \mid \mathcal{X} - \{X, Y\})$$

Distributions to Graphs: Pairwise Independencies

- Given distribution P, we would like to construct the minimal I-map H
- Construct H by adding an edge between all pairs X and Y such that

$$P \not\models (X \perp Y \mid \mathcal{X} - \{X, Y\})$$

Theorem

If P is a positive distribution, the Markov network generated according to pairwise independence policy is the unique minimal I-map for P

Distributions to Graphs: Pairwise Independencies

- Given distribution P, we would like to construct the minimal I-map H
- Construct H by adding an edge between all pairs X and Y such that

$$P \not\models (X \perp Y \mid \mathcal{X} - \{X, Y\})$$

Theorem

If P is a positive distribution, the Markov network generated according to pairwise independence policy is the unique minimal I-map for P

Proof.

(i) By construction, $P \models I_p(H) \Rightarrow P \models I(H)$ (i.e., H is an I-map for P)
(ii) It is minimal since removing any edge X—Y suggests an invalid pairwise independence $(X \perp Y \mid \mathcal{X} - \{X, Y\})$ $\square$

Distributions to Graphs: Markov Blanket

Definition (Markov Blanket (Distribution))

The **Markov blanket** $\mathsf{MB}_P(X)$ of X in a distribution P is the minimal set $\boldsymbol{U}$ such that for $X \notin \boldsymbol{U}$

$$\{X \perp (\boldsymbol{X} - \{X\} - \boldsymbol{U}) \,|\, \boldsymbol{U}\} \in I(P)$$

- Construct H by adding an edge X—Y for all X and $Y \in \mathsf{MB}_P(X)$

Distributions to Graphs: Markov Blanket

Definition (Markov Blanket (Distribution))

The **Markov blanket** $\mathrm{MB}_P(X)$ of X in a distribution P is the minimal set $\boldsymbol{U}$ such that for $X \notin \boldsymbol{U}$

$$\{X \perp (\boldsymbol{X} - \{X\} - \boldsymbol{U}) \,|\, \boldsymbol{U}\} \in I(P)$$

- Construct H by adding an edge X—Y for all X and $Y \in \mathrm{MB}_P(X)$

Theorem

If P is a positive distribution, the Markov network generated according to the Markov blanket policy is the unique minimal I-map for P

Distributions to Graphs: Markov Blanket

Definition (Markov Blanket (Distribution))

The **Markov blanket** $\text{MB}_P(X)$ of X in a distribution P is the minimal set $\boldsymbol{U}$ such that for $X \notin \boldsymbol{U}$

$$\{X \perp (\boldsymbol{X} - \{X\} - \boldsymbol{U}) \,|\, \boldsymbol{U}\} \in I(P)$$

- Construct H by adding an edge X—Y for all X and $Y \in \text{MB}_P(X)$

Theorem

If P is a positive distribution, the Markov network generated according to the Markov blanket policy is the unique minimal I-map for P

Proof.

By construction, $P \models I_l(H) \Rightarrow P \models I(H)$ (i.e., H is an I-map for P) □

Bayesian Networks and Markov Networks

	Bayesian Networks	Markov Networks
local independencies	local Markov	pairwise, Markov blanket
global independencies	d-separation	separation
	CPDs are conditional probabilities	Permit cycles
relative advantages	Joint probability is easy (no partition function)	Evaluating independencies is easy
	More interpretable (causation)	Symmetric