Readings: K&F 4.1, 4.2, 4.3

Undirected Graphical Models I

Lecture 4 – Apr 6, 2011
CSE 515, Statistical Methods, Spring 2011

Instructor: Su-In Lee
University of Washington, Seattle

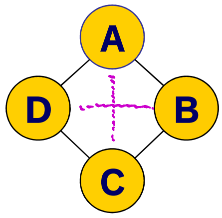
Bayesian Network Representation

- Directed acyclic graph structure
  - Conditional parameterization
  - Independencies in graphs
  - From distribution to BN graphs
- Conditional probability distributions (CPDs)
  - Table
  - Deterministic
  - Context-specific (Tree, Rule CPDs)
  - Independence of causal influence (Noisy OR, GLMs)
  - Continuous variables
  - Hybrid models

The *Misconception* Example

- Four students get together in pairs to work on HWs:
  Alice, **B**ob, **C**harles, **D**ebbie
- Only the following pairs meet: (A&B), (B&C), (C&D), (D&A)
- Let's say that the prof accidentally misspoke in class
  - Each student may subsequently have figured out the problem.
  - In subsequent study pairs, they may transmit this newfound understanding to their partners.
- Consider 4 binary random variables
  - A, B, C, D: whether the student has the misconception or not.
- Independence assumptions?

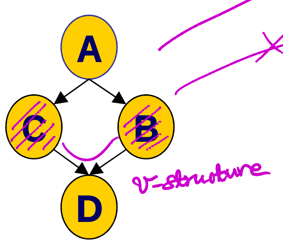
  $(A \perp C \mid B, D)$, $(B \perp D \mid A, C)$

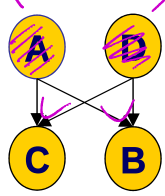


- Can we find the P-map for these?

Reminder: Perfect Maps

- G is a perfect map (P-map) for P if I(P)=I(G)
- Does every distribution have a P-map?
  - No: some structures cannot be represented in a BN
    - **Independencies in P: (A⊥D | B, C) and (B⊥C | A, D)**





(B⊥C | A,D) does not hold

(A⊥D) also holds
unwanted

Representing Dependencies

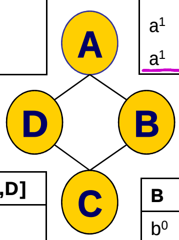
- ($A \perp D \mid B,C$) and ($B \perp C \mid A,D$)
  - Cannot be modeled with a Bayesian network.
  - Can be modeled with an undirected graphical models (Markov networks).

Undirected Graphical Models (Informal)

- **Nodes** correspond to random variables
- **Edges** correspond to direct probabilistic interaction
  - An interaction not mediated by any other variables in the network.
- How to **parameterize**?
- Local factor models are attached to sets of nodes
  - Factor elements are positive
  - Do not have to sum to 1 not a prob.
  - Represent affinities, compatibilities

A	D	π_1[A,C]
a^0	d^0	100
a^0	d^1	1
a^1	d^0	1
a^1	d^1	100

A	B	π_2[A,B]
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



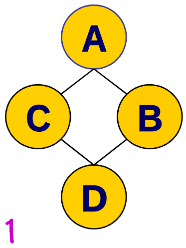
C	D	π_3[C,D]
c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

B	C	π_4[B,C]
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	1000

Undirected Graphical Models (Informal)

- Represents joint distribution
  - Unnormalized factor

$$F(a,b,c,d) = \pi_1[a,b]\pi_2[a,c]\pi_3[b,d]\pi_4[c,d]$$



$F(A=a, B=b, C=c, D=d)$ – factor multiplication

  - Probability $=\pi_1[A=a,B=b]\pi_2[A=a,C=c]\pi_3[B=b,D=d]\pi_4[C=c,D=d]$

$$P(a,b,c,d) = \frac{1}{Z}\pi_1[a,b]\pi_2[a,c]\pi_3[b,d]\pi_4[c,d]$$

$Z \rightarrow$ normalization $\sum_a \cdots \sum_d p(a,b,c,d) = 1$

  - Partition function

$$Z = \sum_{a,b,c,d} \pi_1[a,b]\pi_2[a,c]\pi_3[b,d]\pi_4[c,d]$$

- As undirected graphical models represent joint distributions, they can be used for answering queries.

Undirected Graphical Models Blurb

- Useful when edge directionality cannot be assigned
- Simpler interpretation of structure
  - Simpler inference
  - Simpler independency structure ?
- Harder to learn parameters/structures why?

  e.g. $Z = \underbrace{\sum_{x_1} \cdots \sum_{x_n}} F(x_1, \cdots x_n)$

- We will also see models with combined directed and undirected edges (e.g. conditional random fields)
- Markov networks

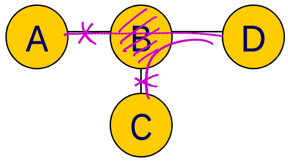
Markov Network Structure

sep$_H$(X;Y|Z)

- Undirected graph H
  - Nodes $X_1,\ldots,X_n$ represent random variables

    BN graph G: $I_l(G)=\{X_i \perp ND(X_i) \mid Pa(X_i)\}$
- H encodes independence assumptions
  - A path X_1-X_2-…-X_k is active if none of the X_i variables along the path are observed
  - **X** and **Y** are separated in H given **Z** if there is no active path between any node x∈**X** and any node y∈**Y** given **Z**
    - Denoted sep$_H$(**X**;**Y**|**Z**)

D ⊥ {A,C} | B



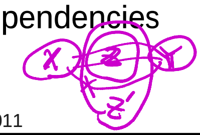
Global independencies associated with H:

I(H) = {(**X**⊥**Y**|**Z**) : sep$_H$(**X**;**Y**|**Z**)}

Relationship with Bayesian Network

- Bayesian network
  - Local independencies → Independence by d-separation (global)
- Markov network $I_L(G)$
  - Global independencies → Local independencies

    separation
- Can all independencies encoded by Markov networks be encoded by Bayesian networks?
  - No, counter example – (A ⊥ B | C,D) and (C ⊥ D | A,B)
- Can all independencies encoded by Bayesian networks be encoded by Markov networks?
  - No, immoral v-structures (explaining away)
- Markov networks encode monotonic independencies
  - If sep$_H$(**X**;**Y**|**Z**) and **Z**⊆**Z'** then sep$_H$(**X**;**Y**|**Z'**)



Markov Network Factors

- A factor (or "potential") is a function from value assignments of a set of random variables **D** to real positive numbers $\Re^+$
  - The set of variables **D** is the scope of the factor

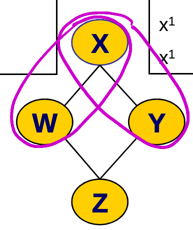
- Factors generalize the notion of CPDs
  - Every CPD is a factor (with additional constraints)

scope of π_1

X	W	π_1[X,W]
x^0	w^0	100
x^0	w^1	1
x^1	w^0	1
x^1	w^1	100

scope of π_2

X	Y	π_2[X,Y]
x^0	y^0	30
x^0	y^1	5
x^1	y^0	1
x^1	y^1	10

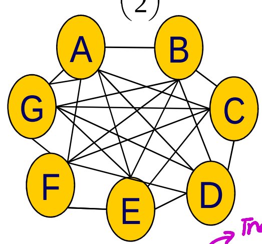


Factors and Joint Distribution

- Can we represent any joint distribution by using only factors that are defined on edges?
  - No! Compare # of parameters
  - Example: n binary RVs
    - Joint distribution has 2^n-1 independent parameters
    - Markov network with edge factors has $4\binom{n}{2}$ parameters

Needed: $2^7-1 = 127!$

Edge parameters: $4\cdot({}_7C_2)=84$



Independencies?

- Factors introduce constraints on joint distribution

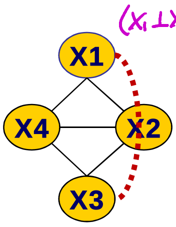
Factors and Graph Structure

- Are there constraints imposed on the network structure H by a factor whose scope is **D**?
  - Hint 1: think of the independencies that must be satisfied
  - Hint 2: generalize from the basic case of |**D**|=2

The induced subgraph over **D** must be a clique (fully connected)

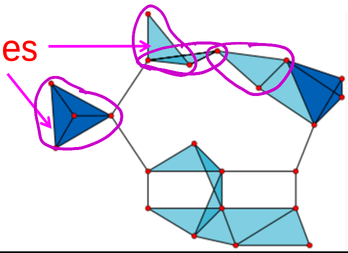
Why? otherwise two unconnected variables may be independent by blocking the active path between them, contradicting the direct dependency between them in the factor over **D**

$(X_1 \perp X_3 \mid \text{something})$



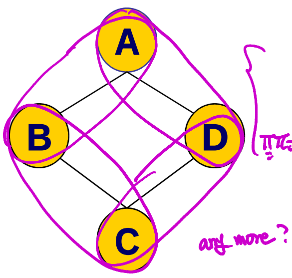
X1,X2,X3,X4	D[x1,x2,x3,x4]
(F,F,F,F)	100
(F,F,F,T)	5
(F,F,TF)	3
(F,F,T,T)	100
⋮	⋮

cliques



Markov Network Factors: Examples

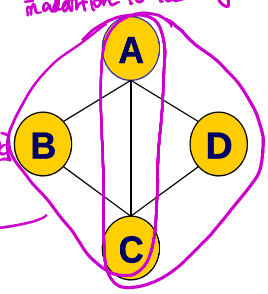
in addition to the edge cliques..



$\prod_i \pi_i[x_i]$

$\prod_i \pi_i[x_i] \prod_{ij} \pi_{ij}[x_i, x_j]$

$\prod_i \pi_i[x_i] \prod_{ij} \pi_{ij}[x_i, x_j] \prod_{ijk} \pi_{ijk}[x_i, x_j, x_k]$



Maximal cliques

- {A,B}
- {B,C}
- {C,D}
- {A,D}

Maximal cliques

- {A,B,C}
- {A,C,D}

Markov Network Distribution

- A distribution P factorizes over H if it has:
  - A set of subsets $\mathbf{D}_1,\ldots\mathbf{D}_m$ where each $\mathbf{D}_i$ is a complete (fully connected) subgraph in H
    clique
  - Factors $\pi_1[\mathbf{D}_1],\ldots,\pi_m[\mathbf{D}_m]$ such that

$$P(X_1,\ldots,X_n)=\frac{1}{Z}f(X_1,\ldots,X_n)=\frac{1}{Z}\prod\pi_i[\mathbf{D}_i]$$

where un-normalized factor: $f(X_1,\ldots,X_n)=\prod\pi_i[\mathbf{D}_i]$

$$Z=\sum_{X_1,\ldots,X_n}f(X_1,\ldots,X_n)=\sum_{X_1,\ldots,X_n}\prod\pi_i[\mathbf{D}_i]$$

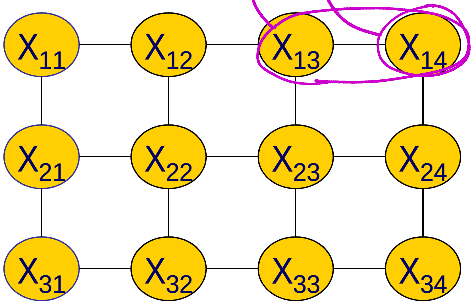
  - Z is called the partition function
  - P is also called a Gibbs distribution over H

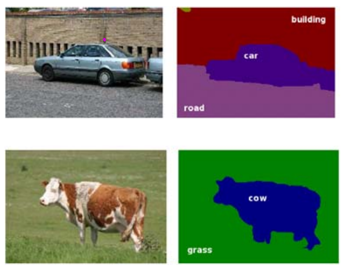
Pairwise Markov Networks

- A **pairwise Markov network** over a graph H has:
  - A set of **node potentials** $\{\pi[X_i]:i=1,\ldots n\}$
  - A set of **edge potentials** $\{\pi[X_i,X_j]: X_i,X_j \in H\}$
  - Example:

neighboring pixels

prior. prob.





Logarithmic Representation

- We represent energy potentials by applying a log transformation to the original potentials
  - $\pi[\mathbf{D}] = \exp(-\varepsilon[\mathbf{D}])$ where $\varepsilon[\mathbf{D}] = -\ln \pi[\mathbf{D}]$
  - Any Markov network parameterized with factors can be converted to a logarithmic representation
  - The log-transformed potentials can take on any real value
  - The joint distribution decomposes as

$$P(X_1,\ldots,X_n) = \frac{1}{Z}\exp\left[-\sum_{i=1}^{m}\varepsilon_i[\mathbf{D}_i]\right]$$

Log P(**X**) is a linear function. of $\varepsilon_i[D_i]$

I-Maps and Factorization

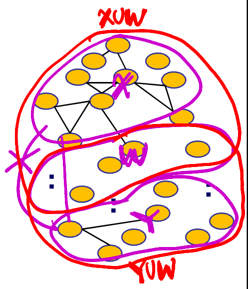
- Independency mappings (I-map)
  - I(P) – set of independencies ($\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$) in P
  - I-map – independencies by a graph is a subset of I(P)

- Bayesian Networks
  - Factorization and reverse factorization theorems
    - G is an I-map of P iff P factorizes as $P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$

- Markov Networks
  - Factorization and reverse factorization theorems
    - H is an I-map of P iff P factorizes as $P(X_1,\ldots,X_n) = \frac{1}{Z}\prod \pi_i[\mathbf{D}_i]$ proof?

Reverse Factorization

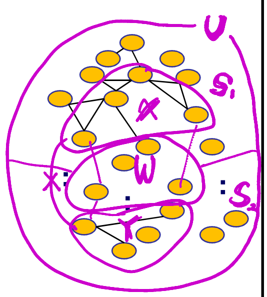
- $P(X_1,...,X_n)=\frac{1}{Z}\prod \pi_i[\mathbf{D}_i] \rightarrow$ H is an I-map of P



- Proof:
  - Let **X**,**Y**,**W** be any three **disjoint** sets of variables such that **W** separates **X** and **Y** in H
  - We need to show (**X**⊥**Y**|**W**)∈I(P)
  - Case 1: **X**∪**Y**∪**W**=U (all variables)
    - As **W** separates **X** and **Y** there are no direct edges between **X** and **Y**
      → any clique in H is fully contained in **X**∪**W** or **Y**∪**W**
    - Let I_X be cliques in **X**∪**W** and I_Y be cliques in **Y**∪**W** (not in I_X)
      → $P(X_1,...,X_n)=\frac{1}{Z}\prod_{i\in I_X}\pi_i[\mathbf{D}_i]\prod_{i\in I_Y}\pi_i[\mathbf{D}_i]=\frac{1}{Z}\overset{\text{no } Y}{f(\mathbf{X},\mathbf{W})}\overset{\text{no } X}{g(\mathbf{Y},\mathbf{W})}$
      → (**X**⊥**Y**|**W**)∈I(P)

Reverse Factorization

- $P(X_1,...,X_n)=\frac{1}{Z}\prod \pi_i[\mathbf{D}_i] \rightarrow$ H is an I-map of P



- Proof:
  - Let **X**,**Y**,**W** be any three **disjoint** sets of variables such that **W** separates **X** and **Y** in H
  - We need to show (**X**⊥**Y**|**W**)∈I(P)
  - Case 2: **X**∪**Y**∪**W**⊂U (all variables)
    - Let **S**=U-(**X**∪**Y**∪**W**)
    - **S** can be partitioned into two disjoint sets $\mathbf{S}_1$ and $\mathbf{S}_2$ such that **W** separates $\mathbf{X}\cup\mathbf{S}_1$ and $\mathbf{Y}\cup\mathbf{S}_2$ in H
    - From case 1, we can derive $(\mathbf{X},\mathbf{S}_1\perp\mathbf{Y},\mathbf{S}_2|\mathbf{W})\in I(P)$
    - From decomposition of independencies
      → (**X**⊥**Y**|**W**)∈I(P)

Factorization

- If H is an I-map of P then $P(X_1,...,X_n) = \frac{1}{Z}\prod \pi_i[\boldsymbol{D}_i]$
- Holds only for positive distributions P
  - Hammerly-Clifford theorem
- Defer proof

Relationship with Bayesian Network

- Bayesian Networks local → global
  - **Semantics** defined via **local** independencies $I_L(G)$.
  - Global independencies induced by d-separation
  - Local and global independencies equivalent since one implies the other
- Markov Networks global $\overset{?}{\to}$ local
  - **Semantics** defined via **global** separation property I(H)
  - Can we define the induced local independencies?
    - We show two definitions (call them "Local Markov assumptions")
    - All three definitions (global and two local) are equivalent only for positive distributions P

Pairwise Independencies

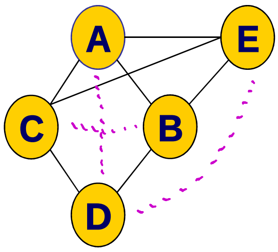
- Every pair of disconnected nodes are separated given all other nodes in the network
- Formally: $I_P(H) = \{ (X \perp Y \mid U - \{X,Y\}) : X—Y \notin H\}$

Example:

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

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

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



Local Independencies

- Every node is independent of all other nodes given <u>its immediate neighboring nodes</u> in the network

  Markov blank of X, $MB_H(X)$

- Formally: $I_L(H) = \{ (X \perp U-\{X\}-MB_H(X) \mid MB_H(X)) : X \in H\}$

Example:

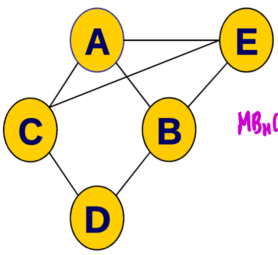
$(A \perp D \mid B,C,E)$

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

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

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

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



$MB_H(A) = \{B, C, E\}$

Relationship Between Properties

- Let I(H) be the global separation independencies
- Let I_L(H) be the local (Markov blanket) independencies
- Let I_P(H) be the pairwise independencies

- For **any** distribution P:
  - I(H) → I_L(H)
    - The assertion in I_L(H), that a node is independent of all other nodes given its neighbors, is part of the separation independencies since there is no active path between a node and its non-neighbors given its neighbors
  - I_L(H) → I_P(H)
    - Follows from the monotonicity of independencies in Markov networks (if $(\mathbf{X}\perp\mathbf{Y}|\mathbf{Z})$ and $\mathbf{Z}\subseteq\mathbf{Z'}$ then $(\mathbf{X}\perp\mathbf{Y}|\mathbf{Z'})$)

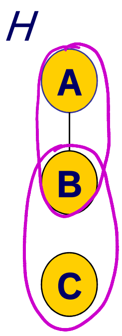
Relationship Between Properties

- Let I(H) be the global separation independencies
- Let I_L(H) be the local (Markov blanket) independencies
- Let I_P(H) be the pairwise independencies

- For any positive distribution P:
  - I_P(H) → I(H)
    - Proof relies on intersection property for probabilities $(\mathbf{X}\perp\mathbf{Y}|\mathbf{Z},\mathbf{W})$ and $(\mathbf{X}\perp\mathbf{W}|\mathbf{Z},\mathbf{Y}) \rightarrow (\mathbf{X}\perp\mathbf{Y},\mathbf{W}|\mathbf{Z})$ which holds in general only for positive distributions
    - **Details on the textbook**
  - Thus, for positive distributions
    - I(H) ↔ I_L(H) ↔ I_P(H)
  - How about a non-positive distribution?

The Need for Positive Distribution

- Let P satisfy
  - A is uniformly distributed
  - A=B=C
- P satisfies $I_P(H)$
  - $(B \perp C|A)$, $(A \perp C|B)$
    (since each variable determines all others)
- P does not satisfy $I_L(H)$
  - $(C \perp A,B)$ needs to hold according to $I_L(H)$ but does not hold in the distribution



Constructing Markov Network for P

- Goal: Given a distribution, we want to construct a Markov network which is an I-map of P
- Complete (fully connected) graphs will satisfy but are not interesting
- Minimal I-maps: A graph G is a minimal I-Map for P if:
  - G is an I-map for P
  - Removing any edge from G renders it not an I-map
- Goal: construct a graph which is a minimal I-map of P

Constructing Markov Network for P

- If P is a positive distribution, then $I(H) \leftrightarrow I_L(H) \leftrightarrow I_P(H)$
  - Thus, sufficient to construct a network that satisfies $I_P(H)$
- Construction algorithm
  - For every (X,Y) add edge if $(X \perp Y | \mathcal{U}\text{-}\{X,Y\})$ does not hold in P
- Theorem: network is minimal and unique I-map
  - Proof:
    - **I-map** follows since $I_P(H)$ by construction and $I(H)$ by equivalence
    - **Minimality** follows since deleting an edge implies $(X \perp Y | \mathcal{U}\text{-}\{X,Y\})$ But, we know by construction that this does not hold in P since we added the edge in the construction process
    - **Uniqueness** follows since any other I-map has at least these edges and to be minimal cannot have additional edges

Summary: Markov Network Representation

- Independencies in graph H
  - Global independencies $I(H) = \{(X \perp Y | Z) : sep_H(X;Y|Z)\}$
  - Local independencies $I_L(H) = \{ (X \perp U\text{-}\{X\}\text{-}MB_H(X) \mid MB_H(X)) : X \in H\}$
  - Pairwise independencies $I_P(H) = \{ (X \perp Y \mid U - \{X,Y\}) : X—Y \notin H\}$
  - For any positive distribution P, they are equivalent.
- (Reverse) factorization theorem: I-map $\leftrightarrow$ factorization
- Markov network factors
  - Has to encompass cliques
  - Maximal cliques, edge potentials
- Pairwise Markov network
  - Node/ edge potentials
  - Application in vision (image segmentation)
- **What next?**
- Log-linear model
  - Log-transformation of potentials
  - Features instead of factors
- Constructing Markov networks from Bayesian networks
- “Partially” directed graph (e.g. Conditional Random Fields)

Acknowledgement

- These lecture notes were generated based on the slides from Prof Eran Segal.