Lecture 4, Unirected Graphical Models

DS-GA 1005 Inference and Representation, Fall 2023

Yoav Wald

09/27/2023

Today's Plan

- Recap of results on Bayesian networks from last week
- Gibbs distributions and Markov networks
- Existence and uniqueness of Markov Networks
- Comparison of directed and undirected graphical models

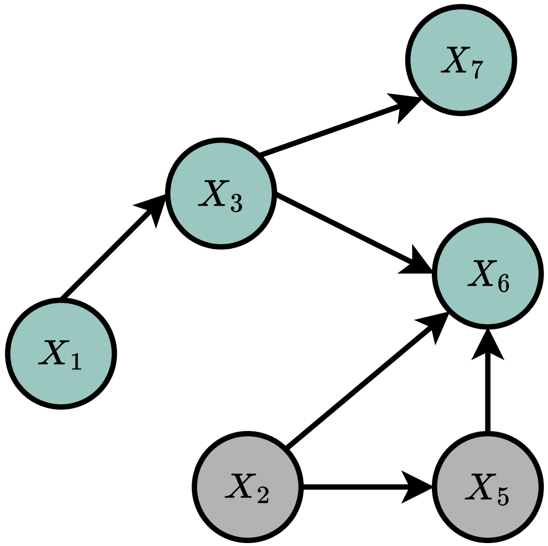
Reminder: Directed Acyclic Graphs (DAGs)

Definition

A directed graph is a data structure $G = (V, E)$ where $E = \{(i, j), i, j \in V\}$ are **ordered** tuples (also $i \to j$). G is acyclic if it has no directed paths from any node $i \in V$ to itself ($i \not\rightsquigarrow i$)

We will usually consider $V = \{X_i\}_{i=1}^d$, where each random variable corresponds to a vertex

- For $i \in V$, we define its parents $Pa(i) = \{j : (j, i) \in E\}$
- **Definition**: P *factorizes over* G if $P(X_1, \ldots, X_d) = \prod_{i=1}^d P(X_i \mid X_{Pa(i)})$



Reminder: D-Separation

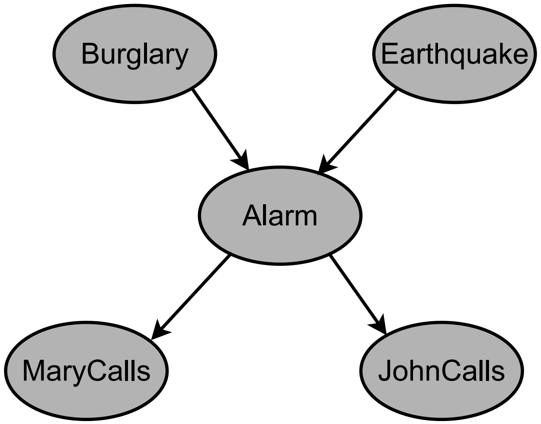
Definition

Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be sets of vertices in G. $\mathbf{X}$ are d-separated from $\mathbf{Y}$ given $\mathbf{Z}$, denoted $d - sep_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ if there is no active trail between any node $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given $\mathbf{Z}$. $\mathcal{I}(G)$ are the set of independencies that correspond to d-separation

$$\mathcal{I}(G) = \{\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \ : \ d - sep_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\}$$

Burglary $\perp$ Earthquake

Burglary $\perp$ JhonCalls | Alarm



Reminder: Factorization $\Leftrightarrow \mathcal{I}(G) \subseteq \mathcal{I}(P)$

Definition

$\mathcal{I}(G)$ are the set of independencies that correspond to d-separation

$$\mathcal{I}(G) = \{\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \ : \ d - sep_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\}$$

Definition (Independence set)

Let P be a distribution over $\mathcal{X} = \{X_1, \ldots, X_d\}$. Then $\mathcal{I}(P)$ is the set of all conditional independence statements of the form $X \perp\!\!\!\perp Z \mid Y$ that hold for P

Theorem (Theorems 3.1, 3.2 and 3.3 in Koller and Friedman)

A distribution P factorizes over G if and only if $\mathcal{I}(G) \subseteq \mathcal{I}(P)$

Reminder: Incompleteness of Representation with Bayesian Network

Some interesting facts about representation with directed graphs:

- Does factorization imply $\mathcal{I}(P) = \mathcal{I}(G)$? Generally, No. But only for measure 0 of distributions that factorize over G
- Can we always find *some* graph such that $\mathcal{I}(P) = \mathcal{I}(G)$? No

$$P(x, y, z) = \begin{cases} 1/12 & x \oplus y \oplus z = 0 \\ 1/6 & x \oplus y \oplus z = 1 \end{cases}$$

- It is simple to show that $X \perp\!\!\!\perp Y$, and from symmetry also that $Y \perp\!\!\!\perp Z$ and $Z \perp\!\!\!\perp X$
- On the other hand, $X \not\perp\!\!\!\perp Y \mid Z$

Conclusion: $\mathcal{I}(G) \neq \mathcal{I}(P) = \{X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z, Y \perp\!\!\!\perp Z\}$ for all G

Today's Plan

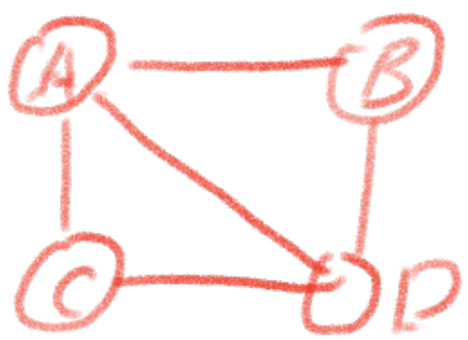
- Recap of results on Bayesian networks from last week
- Gibbs distributions and Markov networks
- Existence and uniqueness of Markov Networks
- Comparison of directed and undirected graphical models

Undirected Graphical Models

Motivation: another graphical representation of distributions that captures different properties from directed models

Definition (Undirected graph and Cliques in a graph)

An undirected graph is a data structure $G = (V, E)$ where $E = \{(i,j), i, j \in V\}$ are **unordered** tuples (also $i - j$).
A **clique** $C \subseteq V$ is a complete subgraph of G. That is, $(i,j) \in E$ for all $i, j \in C$.



(A, B) ✓ clique
(A, C, B) ✗
(A, D, B) ✓

Gibbs Distributions

We consider a distribution $P(X_1, \dots X_d)$ and as for directed models $V = \{X_i\}_{i=1}^d$, we associate each X_i with a vertex in V.

- What is a reasonable way to define *factorization* over an undirected graph G?
- Some notations:
  - For $C \subseteq V$, we denote the set of corresponding random variables by $\mathbf{X}_C = \{X_i : i \in C\}$
  - A non-negative local factor is a function $\Psi_C : Val(\mathbf{X}_c) \to \mathbb{R}_+$, where $Val(\mathbf{X}_c)$ is the set of values that $\mathbf{X}_c$ may take

Gibbs Distributions

Definition

Let $\{\Psi_C\}_{C \in \mathcal{C}}$ be a set of non-negative local factors, where $C \subseteq V$ for any $C \in \mathcal{C}$. P is a Gibbs distribution parameterized by the set if

$$P(X_1, \ldots, X_d) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \Psi_C(\mathbf{X}_C),$$

Partition function

where $Z = \int \prod_{C \in \mathcal{C}} \Psi_C(\mathbf{x}_C) d\mathbf{x}$

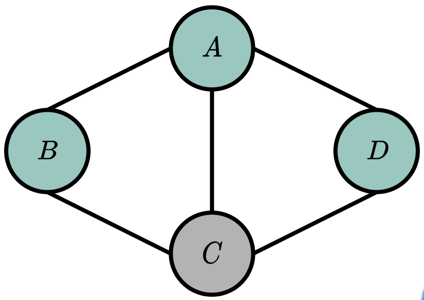
Analogy to Directed Models: we consider decompositions to products of functions over smaller scope

Factorizations and Markov Networks

Definition

A Gibbs distribution P with factors $\{\Psi_C\}_{C \in \mathcal{C}}$ factorizes over an undirected graph G, if each $C \in \mathcal{C}$ is a clique in G. The pair (G, P) is called a Markov network.

- Question: Why did we demand that each $C \in \mathcal{C}$ is a clique?
- Note that there are many possible valid "choices" for $\mathcal{C}$.



Factorization 1: $\Psi(A, B, C)\Psi(A, C, D)$

Maximal clique

Factorization 2: $\Psi(A, B)\Psi(B, C)\Psi(C, D)\Psi(D, A)\Psi(A, C)$

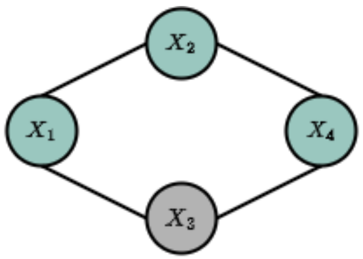
Independencies in Markov Networks

Definition

For an undirected graph G, we denote $sep_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ if all paths from $X \in \mathbf{X}$ to $Y \in \mathbf{Y}$ go through some $Z \in \mathbf{Z}$. The matching set of independencies are

$$\mathcal{I}(G) = \{\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} : sep_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\}$$

- We will prove (soon) that "Factorization $\Rightarrow \mathcal{I}(G) \subseteq \mathcal{I}(P)$"
- First, let us recircle to our question: why did we demand that each $C \in \mathcal{C}$ is a clique?



$$\mathcal{I}(G) = \{X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3,\ X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4\}$$

Why Cliques?

We want to have: Factorization w.r.t $\mathcal{C} \Rightarrow \mathcal{I}(G) \subseteq \mathcal{I}(P)$.

- Consider $(X_i, X_j) \notin E$, then $X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_{[d]\setminus\{i,j\}} \in \mathcal{I}(G)$
- Therefore we should have (where $\mathbf{v}$ is some fixed value)

$$P(X_i, X_j \mid \mathbf{X}_{[d]} \setminus \{i,j\} = \mathbf{v}) = \\ P(X_i \mid \mathbf{X}_{[d]\setminus\{i,j\}} = \mathbf{v}) P(X_j \mid \mathbf{X}_{[d]\setminus\{i,j\}} = \mathbf{v})$$

- Particularly this means that for some functions f, g,

$$P(X_i, X_j \mid \mathbf{X}_{[d]\setminus\{i,j\}} = \mathbf{v}) = f(X_i) g(X_j)$$

Now let us assume $P = \prod_{C \in \mathcal{C}} (\Psi(\mathbf{X}_C))$ and see that no $C \in \mathcal{C}$ satisfies $X_i, X_j \in C$ (Q: why does this mean C is a clique?)

Why Cliques?

$$P(X_i, X_j \mid \mathbf{X}_{[d]\setminus\{i,j\}} = \mathbf{v}) = \frac{P(X_i, X_j, X_{[d]\setminus\{i,j\}})}{\iint \cdots dx_i dx_j}$$

$$\frac{\prod_{C:i,j\in C} \Psi_C(\mathbf{X}_C) \prod_{C:i\in C,j\notin C} \Psi_C(\mathbf{X}_C) \prod_{C:i\notin C,j\in C} \Psi_C(\mathbf{X}_C) \prod_{C:i,j\notin C} \Psi_C(\mathbf{X}_C)}{\int \prod_{C:i,j\in C} \Psi_C(\mathbf{X}_C) \ldots \prod_{C:i,j\notin C} \Psi_C(\mathbf{X}_C) \mathrm{d}x_i \mathrm{d}x_j}$$

$$= \frac{\prod_{C:i,j\in C} \Psi_C(\mathbf{X}_C) \prod_{C:i\in C,j\notin C} \Psi_C(\mathbf{X}_C) \prod_{C:i\notin C,j\in C} \Psi_C(\mathbf{X}_C) \cancel{\prod_{C:i,j\notin C} \Psi_C(\mathbf{X}_C)}}{\int \prod_{C:i,j\in C} \Psi_C(\mathbf{X}_C) \ldots \mathrm{d}x_i \mathrm{d}x_j \cancel{\prod_{C:i,j\notin C} \Psi_C(\mathbf{X}_C)}}$$

Why Cliques?

$$P(X_i, X_j \mid \mathbf{X}_{[d]\setminus\{i,j\}} = \mathbf{v}) =$$

$$= \frac{\underbrace{\prod_{C:i,j\in C} \Psi_C(\mathbf{X}_C)}_{\text{function of both } X_i, X_j} \underbrace{\prod_{C:i\in C, j\notin C} \Psi_C(\mathbf{X}_C)}_{\text{function of } X_i} \underbrace{\prod_{C:i\notin C, j\in C} \Psi_C(\mathbf{X}_C)}_{\text{function of } X_j}}{\underbrace{\int \prod_{C:i,j\in C} \Psi_C(\mathbf{X}_C) \ldots \mathrm{d}x_i \mathrm{d}x_j}_{\text{not a function of } X_i, X_j}}$$

Conclusion: $P(X_i, X_j \mid \mathbf{X}_{[d]\setminus\{i,j\}})$ is a product of $f(X_i)g(X_j)$ only if there are no factors where $i, j \in C$

Independencies in Markov Networks

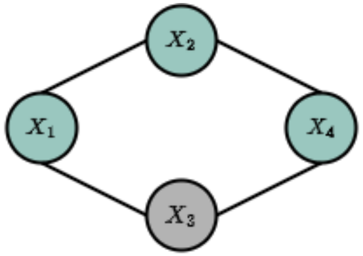
Definition

For an undirected graph G, we denote $sep_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ if all paths from $X \in \mathbf{X}$ to $Y \in \mathbf{Y}$ go through some $Z \in \mathbf{Z}$. The matching set of independencies are

$$\mathcal{I}(G) = \{\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} : sep_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\}$$

- We will prove (soon) that "Factorization $\Rightarrow \mathcal{I}(G) \subseteq \mathcal{I}(P)$"
- Like in Bayesian networks, we want "$\mathcal{I}(G) \subseteq \mathcal{I}(P) \Rightarrow$ Factorization"

  *this will hold with a small caveat



$$\mathcal{I}(G) = \{X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3,\ X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4\}$$

Factorization $\Rightarrow \mathcal{I}(G) \subseteq \mathcal{I}(P)$: Toy Example

- Consider the graph $A - B - C$, where $I(G) = \{A \perp\!\!\!\perp C \mid B\}$
- Let us prove that if P factorizes over G then $\mathcal{I}(G) \subseteq \mathcal{I}(P)$
  - By definition of factorization we have

$$P(A, B, C) = \frac{1}{Z}\Psi(A, B)\Psi(B, C)$$

  - Write down the conditional distributions:

$$P(A = a \mid B = b) = \ldots$$

Factorization $\Rightarrow \mathcal{I}(G) \subseteq \mathcal{I}(P)$: Toy Example

$$
\begin{aligned}
P(A = a \mid B = b) &= \frac{P(a,b)}{P(b)} = \frac{\int \Psi_{AB}(a,b)\Psi_{BC}(b,c)\mathrm{d}c}{\int\int \Psi_{AB}(a,b)\Psi_{BC}(b,c)\mathrm{d}a\mathrm{d}c} \\
&= \frac{\Psi_{AB}(a,b)\cdot\left(\int \cancel{\Psi_{BC}(b,c)}\mathrm{d}c\right)}{\left(\int \Psi_{AB}(a,b)\mathrm{d}a\right)\cdot\left(\int \cancel{\Psi_{BC}(b,c)}\mathrm{d}c\right)} \\
&= \frac{\Psi_{AB}(a,b)}{\int \Psi_{AB}(a,b)\mathrm{d}a}
\end{aligned}
$$

Factorization $\Rightarrow \mathcal{I}(G) \subseteq \mathcal{I}(P)$: Toy Example

- The graph $A - B - C$, where $I(G) = \{A \perp\!\!\!\perp C \mid B\}$
- We prove that if $P(A, B, C) = \frac{1}{Z}\Psi(A, B)\Psi(B, C)$ then $\mathcal{I}(G) \subseteq \mathcal{I}(P)$
- Write down the conditional distributions:

$$P(A = a \mid B = b) = \frac{\Psi_{AB}(a, b)}{\int \Psi_{AB}(a, b)\mathrm{d}a},$$

$$P(C = c \mid B = b) = \frac{\Psi_{BC}(b, c)}{\int \Psi_{BC}(b, c)\mathrm{d}c}$$

- Now from Bayes rule

$$\begin{aligned} P(a, c \mid B = b) &= \frac{\Psi_{AB}(a, b)\Psi_{BC}(b, c)}{\int \Psi_{AB}(a, b)\mathrm{d}a \int \Psi_{BC}(b, c)\mathrm{d}c} \\ &= P(a \mid B = b)P(c \mid B = b) \qquad \square \end{aligned}$$

Factorization $\Rightarrow \mathcal{I}(G) \subseteq \mathcal{I}(P)$: General Case

Theorem (Thm. 4.1, Koller and Friedman)

If P is a Gibbs distribution that factorizes over G then $\mathcal{I}(G) \subseteq \mathcal{I}(P)$

Proof sketch.

- If $sep(\mathbf{X}; \mathbf{Z} \mid \mathbf{Y})$, or in other words $\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{Y} \in \mathcal{I}(G)$, then there are no direct edges between $\mathbf{X}$ and $\mathbf{Z}$
- $\Rightarrow$ All cliques are either in $\mathbf{X} \cup \mathbf{Y}$ or $\mathbf{Z} \cup \mathbf{Y}$
- $\Rightarrow$ P must factorize as $P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \frac{1}{Z} f(\mathbf{X}, \mathbf{Y}) g(\mathbf{Y}, \mathbf{Z})$
- Complete proof along the lines of toy example

$\square$

Independencies in Markov Networks

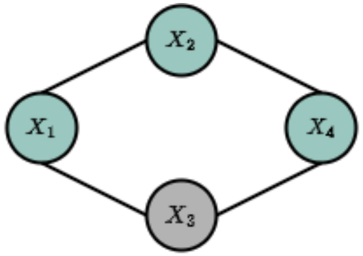
Definition

For an undirected graph G, we denote $sep_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ if all paths from $X \in \mathbf{X}$ to $Y \in \mathbf{Y}$ go through some $Z \in \mathbf{Z}$. The matching set of independencies are

$$\mathcal{I}(G) = \{\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} : sep_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\}$$

- We will prove (soon) that "Factorization $\Rightarrow \mathcal{I}(G) \subseteq \mathcal{I}(P)$"
- Like in Bayesian networks, we want "$\mathcal{I}(G) \subseteq \mathcal{I}(P) \Rightarrow$ Factorization"

  *this will hold with a small caveat



$$\mathcal{I}(G) = \{X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3,\ X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4\}$$

Hammersley Clifford Theorem

Theorem (Hammersley-Clifford)

Let P be a **positive** *distribution over $\mathcal{X}$ and G an undirected graph over $\mathcal{X}$. If $\mathcal{I}(G) \subseteq \mathcal{I}(P)$, then P is a Gibbs distribution w.r.t G*

Proof.

In the next recitation □

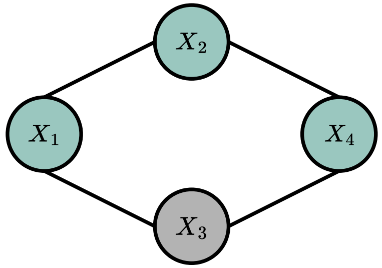
Conclusion: If P is positive then factorization is equivalent to independence

Bayesian Networks vs. Markov Networks

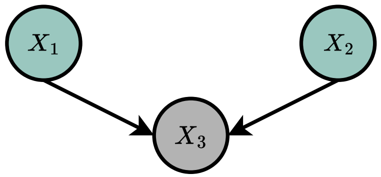
Let us look back at the similarities and differences between the directed and undirected case

	Factorization		Independence
undirected $\circ - \circ$	$P(x_1,\dots,x_d)=\frac{1}{Z}\prod_{c\in C}\psi_c(X_c)$ Gibbs dist., c is clique in G	$\Rightarrow$ $\Leftarrow$ $(p>0)$	Markov Assumption $I(G)= X\perp Z \mid Y$ if Y separates X and Z
directed $\circ \to \circ$	$P(x_1,\dots,x_d)=\prod_i P(x_i \mid X_{Pa(i)})$ Factorizes over G if $Pa(i)$ are i's parents in G.	$\Leftarrow$ $\Rightarrow$	d-separation $I(G)= X\perp Z \mid Y$ if X and Z are d-separated given Y

- Do we really need both directed and undirected representations? Yes, because they capture different sets of independencies



$\mathcal{I}(G) = \{X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3,\ X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4\}$



$\mathcal{I}(G) = \{X_1 \perp\!\!\!\perp X_2\}$

Bayesian Networks vs. Markov Networks

- Do we really need both directed and undirected representations? Yes, because they capture different sets of independencies
- How can we interpret factors in Markov networks?
  - In Bayesian networks they were Conditional Probability Distributions, $P(X_i \mid \mathbf{X}_{Pa(i)})$
  - In Markov networks they do not have special meaning, consider $X - Y - Z$

$$
\begin{aligned}
P(X, Y, Z) &= \overbrace{P(X \mid Y) P(Y)}^{\psi(X,Y)} \overbrace{P(Z \mid Y)}^{\psi(Z,Y)} \\
&= P(X \mid Y) P(Y) P(Z \mid Y) \\
&= P(X \mid Y) \sqrt{P(Y)} \sqrt{P(Y)} P(Z \mid Y)
\end{aligned}
$$

Bayesian Networks vs. Markov Networks

- Do we really need both directed and undirected representations? Yes, because they capture different sets of independencies
- How can we interpret factors in Markov networks? No special interpretation
- Main distinction between the networks are v-structures (or colliders)
  - Directed models are good for modelling "explaining away"
  - Undirected models are more suitable to capture symmetric relations

Bayesian Networks vs. Markov Networks

- Do we really need both directed and undirected representations? Yes, because they capture different sets of independencies
- How can we interpret factors in Markov networks? No special interpretation
- Main distinction between the networks are v-structures (or colliders)
- *Note:* our discussion was mainly around discrete variables, under some assumptions similar results hold for continuous domains

Going Forward: Inference and Learning

What's up ahead

- Representation is nice, but how do we use it?
- Inference (next two weeks): Assume we have the CPDs of a Bayesian network, or factors of a markov network. How do we answer queries?
  - $P(\text{Asthma} \mid Symptoms)$
- Learning (in 2 weeks): Given raw data $\{\mathbf{x}_i\}_{i=1}^m$ sampled from P, how do we estimate the parameters for a model $P_\theta(\mathbf{x})$?

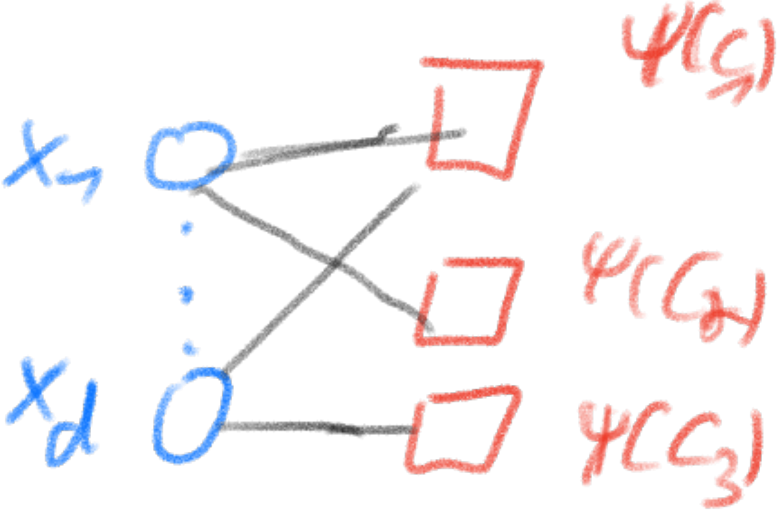
Factor Graphs

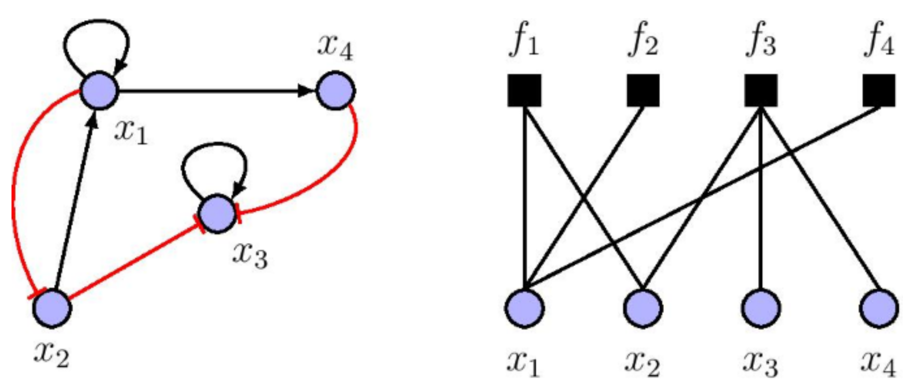
Do we need different inference algorithms for directed and undirected graphs?

Definition

A factor graph is a bipartite graph where

- Nodes correspond to both variables and potential factors $\{\Psi_C\}_{C \in \mathcal{C}}$
- $E = \{(i, C) : i \in C\}$, that is we draw an edge when a variable X_i appears in factor C





(a) (b)

Factor Graphs

- Both directed and undirected models can be embedded in factor graphs and factorized as $P(X_1, \ldots, X_d) = \prod_{C \in \mathcal{C}} \Phi(\mathbf{X}_C)$. Size of factors does not change.
- Next week: The first inference algorithm we will learn!
  *works on factor graphs; applicable to both directed and undirected models