Graphical Models

Undirected Models

Siamak Ravanbakhsh  
Winter 2018
Learning objectives

Markov networks:
- independence assumptions
- factorization
- representations:
  - factor-graph
  - log-linear models
Challenge

Given the following set of CIs draw their DAG

\[ \mathcal{I}(P) = \{ (A \perp C \mid B, D), (D \perp B \mid A, C) \} \]
Challenge

Given the following set of CIs draw their DAG

\(\mathcal{I}(P) = \{(A \perp C \mid B, D), (D \perp B \mid A, C)\}\)
Challenge

Given the following set of CIs draw their DAG

\[ \mathcal{I}(P) = \{(A \perp C \mid B, D), (D \perp B \mid A, C)\} \]

a DAG cannot be a P-map for P
an undirected model can!
Challenge

Given the following set of CIs draw their DAG

\[ \mathcal{I}(P) = \{(A \perp C \mid B, D), (D \perp B \mid A, C)\} \]

a DAG cannot be a P-map for P
an undirected model can!
Motivation

Statistical physics: **Ising model** of ferromagnetism

Image: https://web.stanford.edu/~peastman/statmech/phasetransitions.html

CIs are naturally expressed using an undirected model
Motivation

Social sciences

Cls are naturally expressed using an undirected model
Motivation

Combinatorial problems

CIs are naturally expressed using an undirected model.

Graph coloring
**Factorization in Markov networks**

\[
P(A, B, C, D) = \frac{1}{Z} \phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(A, D)
\]

\[
Z = \sum_{a, b, c, d} \phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(a, d)
\]

is a normalization constant (*partition function*)

\[
\phi_1 : \text{Val}(A, B) \to [0, +\infty) \text{ is called a factor (potential)}
\]
MRF; Conditional Independencies

\[ P(A, B, C, D) = \frac{1}{Z} \phi_1(A, B) \phi_2(B, C) \phi_3(C, D) \phi_4(A, D) \]

\[ P \models (B \perp D \mid A, C) \]

\[ P(A, B, C, D) = \frac{1}{Z} \phi_1(A, B) \phi_2(A, D) \phi_3(C, D) \phi_4(B, C) \]

\[ P \models (A \perp C \mid B, D) \]
Product of factors

\[ P(A, B, C, D) = \frac{1}{Z} \phi_1(A, B) \phi_2(B, C) \phi_3(C, D) \phi_4(A, D) \]

\[ \psi(A, B, C) : \text{Val}(A, B, C) \rightarrow \mathbb{R}^+ \]

\[ \text{Val}(A) \times \text{Val}(B) \times \text{Val}(C) \quad \text{similar to a 3D tensor} \]
Q: Do factors represent marginals?

**Simplified example:** \( P(A, B, C) = \frac{1}{Z} \phi_1(A, B)\phi_2(B, C) \)

\[
\begin{array}{ccc}
  a^1 & b^1 & c^1 \\
  a^1 & b^2 & c^2 \\
  a^2 & b^1 & c^1 \\
  a^2 & b^2 & c^2 \\
  a^3 & b^1 & c^1 \\
  a^3 & b^2 & c^2 \\
\end{array}
\]

\[
\begin{array}{ccc}
  a^1 & b^1 & 0.5 \\
  a^1 & b^2 & 0.8 \\
  a^2 & b^1 & 0.1 \\
  a^2 & b^2 & 0 \\
  a^3 & b^1 & 0.3 \\
  a^3 & b^2 & 0.9 \\
\end{array}
\]

\[
\begin{array}{ccc}
  a^1 & b^1 & c^1 \\
  a^1 & b^2 & c^2 \\
  a^2 & b^1 & c^1 \\
  a^2 & b^2 & c^2 \\
  a^3 & b^1 & c^1 \\
  a^3 & b^2 & c^2 \\
\end{array}
\]

\[
\begin{array}{ccc}
  a^1 & b^1 & 0.5 \cdot 0.5 = 0.25 \\
  a^1 & b^2 & 0.5 \cdot 0.7 = 0.35 \\
  a^1 & b^2 & 0.8 \cdot 0.1 = 0.08 \\
  a^1 & b^2 & 0.8 \cdot 0.2 = 0.16 \\
  a^2 & b^1 & 0.1 \cdot 0.5 = 0.05 \\
  a^2 & b^2 & 0.1 \cdot 0.7 = 0.07 \\
  a^3 & b^1 & 0.3 \cdot 0.5 = 0.15 \\
  a^3 & b^1 & 0.3 \cdot 0.7 = 0.21 \\
  a^3 & b^2 & 0.9 \cdot 0.1 = 0.09 \\
  a^3 & b^2 & 0.9 \cdot 0.2 = 0.18 \\
\end{array}
\]

\[
P(A, B, C) \times Z = .25 + .35 + \ldots = 1.55
\]

**Marginal probabilities:**

\[
P(a^1, b^1) = (.25 + .35)/Z \approx .38
\]

\[
P(a^1, b^2) = (.08 + .16)/Z \approx .15
\]

**Compare to** \( \phi_1 \)

\[
\phi_1(a^1, b^1) = .5
\]

\[
\phi_1(a^1, b^2) = .8
\]
Factorization: general form

\[ P(X) = \frac{1}{Z} \prod_k \phi_k(D_k) \]

\textbf{Gibbs distribution}

\( P \) factorizes over the cliques

Can always convert to factorization over \textit{maximal cliques}
Factorization: general form

$P$ factorizes over cliques

\[ P(X) = \frac{1}{Z} \prod_k \phi_k(D_k) \]

Rewrite as factorization over maximal cliques

- original form of $P$
  \[ P(A, B, C, D) = \phi_1(A, B)\phi_2(A, D)\phi_3(B, D)\phi_4(C, D)\phi_5(B, C) \]

- factorized over cliques
  \[ P(A, B, C, D) = \psi_1(A, B, C)\psi_2(B, C, D) \]
Factorized form: directed vs undirected

Markov Networks:

\[ P(X) = \frac{1}{Z} \prod_k \phi_k(D_k) \]

Bayesian Networks:

\[ P(X) = \prod_k P(X_i \mid Pa_{X_i}) \]

- No partition function
- Each factor is a cond. distribution
- One factor per variable
Conditioning on the evidence

given \( P(X) \propto \prod_k \phi_k(D_k) \), how to obtain \( P(X \mid U = u) \)?

fix the evidence in the relevant factors \( P(X \mid U = u) \propto \prod_k \phi_k[U = u] \)

\[
\begin{array}{|c|c|c|c|}
\hline
a^1 & b^1 & c^1 & 0.5 \cdot 0.5 = 0.25 \\
\hline
a^1 & b^1 & c^2 & 0.5 \cdot 0.7 = 0.35 \\
\hline
a^1 & b^2 & c^1 & 0.8 \cdot 0.1 = 0.08 \\
\hline
a^1 & b^2 & c^2 & 0.8 \cdot 0.2 = 0.16 \\
\hline
a^2 & b^1 & c^1 & 0.1 \cdot 0.5 = 0.05 \\
\hline
a^2 & b^1 & c^2 & 0.1 \cdot 0.7 = 0.07 \\
\hline
a^2 & b^2 & c^1 & 0 \cdot 0.1 = 0 \\
\hline
a^2 & b^2 & c^2 & 0 \cdot 0.2 = 0 \\
\hline
a^3 & b^1 & c^1 & 0.3 \cdot 0.5 = 0.15 \\
\hline
a^3 & b^1 & c^2 & 0.3 \cdot 0.7 = 0.21 \\
\hline
a^3 & b^2 & c^1 & 0.9 \cdot 0.1 = 0.09 \\
\hline
a^3 & b^2 & c^2 & 0.9 \cdot 0.2 = 0.18 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|}
\hline
a^1 & b^1 & c^1 & 0.25 \\
\hline
a^1 & b^2 & c^1 & 0.08 \\
\hline
a^2 & b^1 & c^1 & 0.05 \\
\hline
a^2 & b^1 & c^2 & 0 \\
\hline
a^3 & b^1 & c^1 & 0.15 \\
\hline
a^3 & b^2 & c^1 & 0.09 \\
\hline
\end{array}
\]

\[\phi_k[A, B, C] \text{ conditioned on } C = c^1 \]

\[\phi_k[C = c] \]
Conditioning on the evidence
effect on the graphical model

- cannot create new dependencies
- compare this to colliders in Bayes-nets
Pairwise conditional independencies

Non-adjacent nodes are independent given everything else

\( X \perp Y \mid \mathcal{X} - \{X, Y\} \)
Local conditional Independencies

\[ MB^\mathcal{H}(X) : \text{Markov blanket} \text{ of node } X \text{ in graph } \mathcal{H} \]

\[ X \perp \mathcal{X} - X - MB^\mathcal{H}(X) \mid MB^\mathcal{H} \]

Given its Markov blanket \( X \) is independent of every other variable
**Local conditional Independencies**

\( MB^H(X) \) : **Markov blanket** of \( X \) in graph \( H \)

\[
X \perp \mathcal{X} - X - MB^H(X) \mid MB^H
\]

\( MB^G(X) \) : **Markov blanket** of \( X \) in DAG \( G \)

- Parents
- Children
- Parents of children

\[
X \perp \mathcal{X} - X - MB^G(X) \mid MB^G
\]
Global conditional Independencies

\( X \perp Y \mid Z \) iff every path between \( X \) and \( Y \) is blocked by \( Z \)

much simpler than D-separation
Relationship between the three

pairwise $\mathcal{I}_p$  $\iff$  local $\mathcal{I}_\ell$  $\iff$  global $\mathcal{I}$

\[
\begin{align*}
\begin{array}{c}
\text{pairwise $\mathcal{I}_p$} \\
\begin{array}{c}
\text{local $\mathcal{I}_\ell$} \\
\begin{array}{c}
\text{global $\mathcal{I}$}
\end{array}
\end{array}
\end{array}
\end{align*}
\]

\[
\begin{align*}
\text{(X \perp Y \mid Z')} \\
\text{(X \perp Y \mid Z)}
\end{align*}
\]
Relationship between the three

\[ \text{pairwise } \mathcal{I}_p \quad \leftarrow \quad \text{local } \mathcal{I}_\ell \quad \leftarrow \quad \text{global } \mathcal{I} \]

\[ (X \perp Y \mid Z') \quad \Rightarrow \quad \mathcal{I} \]
Factorization & independence

Recall this relationship in **Bayesian Networks**:  
- Factorization according to a DAG  
- Local & global CIs

Is it similar for **Markov Networks**?  
- Factorization according to an *undirected graph*  
- Pairwise, local & global CIs

(same family of distributions)
Factorization & Independence

Is it similar for Markov Networks?

- **Factorization** according to an *undirected graph*
- **Pairwise, local & global** CIs

⚡ **Short answer:**

- for positive distributions they are equivalent
Factorization $\Rightarrow$ CI

given $P(X) \propto \prod_k \phi_k(C_k)$ does \textit{local} CI hold?
Factorization $\Rightarrow$ CI

given $P(X) \propto \prod_k \phi_k(C_k)$ does local CI hold?

proof
Factorization $\Rightarrow$ CI

given $P(\mathbf{X}) \propto \prod_k \phi_k(\mathbf{C}_k)$ does local CI hold?

\textbf{proof}

$P(\mathbf{X}) \propto \prod_k \phi_k(\mathbf{C}_k) = \prod_{\mathbf{C}_k \in MB(\mathbf{X}_i)} \phi_k(\mathbf{C}_k) \prod_{\mathbf{C}_k \notin MB(\mathbf{X}_i)} \phi_k(\mathbf{C}_k)$

$= f(\mathbf{X}_i, MB(\mathbf{X}_i)) g(\mathbf{X} - \mathbf{X}_i) \Rightarrow$

$\mathbf{X}_i \perp \mathbf{X} - MB^H(\mathbf{X}_i) - \mathbf{X}_i \mid MB^H(\mathbf{X}_i)$
Hammersely-Clifford theorem:

If $P$ is strictly positive satisfying CI $\mathcal{I}(\mathcal{H})$ then $P$ factorizes over $\mathcal{H}$

proof: needs canonical parametrization
Parametrization: redundancy

is this representation of P unique?

\[ P(A, B, C) \propto \phi_1(A, B)\phi_2(B, C)\phi_3(C, A) \]
Parametrization: redundancy

is this representation of P unique?

\[ P(A, B, C) = \frac{1}{2} \phi_1(A, B) \phi_2(B, C) \phi_3(C, A) \]

multiplying all factors by a constant only affects \( Z \)
Parametrization: redundancy

is this representation of P unique?

\[ P(A, B, C) = \frac{1}{2} \phi_1(A, B)\phi_2(B, C)\phi_3(C, A) \]

use the logarithmic form

\[ P(A, B, C) = \frac{1}{Z} 2(\psi_1(A, B) + \psi_2(B, C) + \psi_3(C, A)) \]
Parametrization: redundancy

Is this representation of $P$ unique?

$$P(A, B, C) = \frac{1}{2} \phi_1(A, B) \phi_2(B, C) \phi_3(C, A) \phi_4(B) \phi_5(A) \phi_6(C)$$

use the logarithmic form

$$P(A, B, C') = \frac{1}{Z} 2(\psi_1(A,B) + \psi_2(B,C) + \psi_3(C,A))$$

simplify using local potentials
Parametrization: redundancy

is this representation of P unique?

\[ P(A, B, C) = \frac{1}{Z} \phi_1(A, B)\phi_2(B, C)\phi_3(C, A)\phi_5(A)\phi_6(C) \]

use the logarithmic form

\[ P(A, B, C') = \frac{1}{Z} 2(\psi_1(A,B)+\psi_2(B,C)+\psi_3(C,A)) \]

simplify using local potentials

log-values
**Parametrization:** example (Ising model)

**Ising model:** \( \text{Val}(X_i) = \{-1, +1\} \quad p(x) = \frac{1}{Z(t)} \exp \left( -\frac{1}{t} \left( \sum_i h_i x_i + \frac{1}{2} \sum_{i,j \in \mathcal{E}} x_i J_{ij} x_j \right) \right) \)

*can represent all positive, pairwise Markov networks over the binary domain*

![Image](https://web.stanford.edu/~peastman/statmech/phase_transitions.html)
Parametrization: example (Boltzmann machine)

**Boltzmann machine:** $Val(X_i) = \{0, 1\}$

$$p(x) = \frac{1}{Z} \exp \left( - \sum_i b_i x_i - \frac{1}{2} \sum_{i,j \in E} x_i W_{ij} x_j \right)$$
Parametrization: **log-linear model**

for a positive distribution:

\[ P(X) \propto \prod_k \phi_k(D_k) = \exp(- \sum_k \psi_k(D_k)) \]

\[ \underbrace{\text{energy}}_{\text{energy}} - \log(\phi_k(D_k)) \]
Parametrization: log-linear model

for a positive distribution:

\[ P(X) \propto \prod_k \phi_k(D_k) = \exp(-\sum_k \psi_k(D_k)) \]

linearly parameterize it:

\[ P_w(X) \propto \exp(-\sum_k w_k f_k(D_k)) \]
Parametrization: **log-linear model**

for a positive distribution:

\[ P(X) \propto \prod_k \phi_k(D_k) = \exp\left( - \sum_k \psi_k(D_k) \right) \]

linearly parameterize it:

\[ P_w(X) \propto \exp\left( - \sum_k w_k f_k(D_k) \right) \]

revisit in the exponential family
Parametrization: log-linear model

features in discrete distributions:

\[ P_w(X) \propto \exp\left( - \sum_k w_k f_k(D_k) \right) \]

\[ f_{1,1}(A, B) = \mathbb{I}(A = a^1, B = b^1) \]
Parametrization: log-linear model

\[ P_w(X) \propto \exp\left( -\sum_k w_k f_k(D_k) \right) \]

\[
\begin{align*}
  f_{1,1}(A, B) &= \mathbb{I}(A = a^1, B = b^1) \\
  f_{0,1}(A, B) &= \mathbb{I}(A = a^0, B = b^1) \\
  f_{1,0}(A, B) &= \mathbb{I}(A = a^1, B = b^0) \\
  f_{0,0}(A, B) &= \mathbb{I}(A = a^0, B = b^0)
\end{align*}
\]
**Parametrization: log-linear model**

\[ P_w(X) \propto \exp\left( -\sum_k w_k f_k(D_k) \right) \]

\[
\begin{align*}
  f_{1,1}(A, B) &= \mathbb{I}(A = a^1, B = b^1) \\
  f_{0,1}(A, B) &= \mathbb{I}(A = a^0, B = b^1) \\
  f_{1,0}(A, B) &= \mathbb{I}(A = a^1, B = b^0) \\
  f_{0,0}(A, B) &= \mathbb{I}(A = a^0, B = b^0)
\end{align*}
\]

*Overparameterized model: \( \{w_k\} \rightarrow P_w \) is not one-to-one*
Parametrization: **log-linear model**

\[ P_w(X) \propto \exp(- \sum_k w_k f_k(D_k)) \]

Redundant \( \equiv \) linearly dependent features

\[ \sum_k \alpha_k f_k(D) = \alpha \quad \forall D \]

\[ P_w(X) \propto \exp(- \sum_k w_k f_k(D_k)) \propto \exp(- \sum_k (w_k + \alpha_k) f_k(D_k)) \propto P_{w+\alpha}(X) \]
Parametrization: **log-linear model**

\[
P_w(X) \propto \exp \left( - \sum_k w_k f_k(D_k) \right)
\]

\[
f_{1,1}(A, B) = \mathbb{I}(A = a^1, B = b^1)
\]
\[
f_{0,1}(A, B) = \mathbb{I}(A = a^0, B = b^1)
\]
\[
f_{1,0}(A, B) = \mathbb{I}(A = a^1, B = b^0)
\]
\[
f_{0,0}(A, B) = \mathbb{I}(A = a^0, B = b^0)
\]

**Linear dependency of features:**

\[
f_{0,0}(A, B) + f_{1,0}(A, B) + f_{0,1}(A, B) + f_{1,1}(A, B) = 1
\]
Markov network representation:

- identifies CIs
- defines the factorized form

- is not fine-grained enough

\[ P(A, B, C) = \phi_1(A, B)\phi_2(B, C)\phi_3(C, A) \]
\[ P(A, B, C) = \phi_1(A, B, C) \]
Parametrization: factor-graph

use a bipartite structure:

- factors (square)
- variables (circle)

\[ P(A, B, C) = \phi_1(A, B)\phi_2(B, C)\phi_3(C, A) \]

\[ P(A, B, C) = \phi(A, B, C) \]
Summary

- similar to directed models:
  - factorization of the probability over cliques
  - set of conditional independencies
    - (pairwise, local, global)

\[ P > 0 \implies \text{same family of dists.} \]
Summary

• similar to directed models:
  - factorization of the probability over cliques
  - set of conditional independencies
    - (pariwise, local, global)

• parametrization
  - redundancy (same dist. different params/factors)
  - log-linear model
  - factor-graph (finer-grained specification of the factors)

\[ P > 0 \implies \text{same family of dists.} \]
Parametrization: canonical form

reparameterize a given Gibbs dist.

\[ P(X) \propto \exp\left(- \sum_k \psi_k(D_k)\right) \]

such that low order interactions are automatically moved to smaller cliques

need to fixed an assignment \( \xi^* = (x_1^*, \ldots, x_n^*) \) e.g., \( \xi^* = (0, \ldots, 0) \)
Mobius inversion lemma

For two functions $f, g : 2^\mathcal{X} \to \mathbb{R}$ defined over all subsets $\mathcal{Z} \subseteq \mathcal{X}$ the following are equivalent:

\[ \forall \mathcal{Z} \subseteq \mathcal{X} \quad f(\mathcal{Z}) = \sum_{S \subseteq \mathcal{Z}} g(S) \]
\[ \forall \mathcal{Z} \subseteq \mathcal{X} \quad g(\mathcal{Z}) = \sum_{S \subseteq \mathcal{Z}} (-1)^{|\mathcal{Z} - S|} f(S) \]
Parametrization: \textit{canonical form}

Given a \textit{fixed an assignment} \( \xi^* = (x_1^*, \ldots, x_n^*) \) \ e.g., \( \xi^* = (0, \ldots, 0) \)

\[
f(x_Z) \triangleq \log P(x_Z, \xi_Z^*) \text{ is defined for all } Z \subseteq \{1, \ldots, N\}
\]

\[
f(x) = \log P(x)
\]
Parametrization: canonical form

Given a fixed assignment $\xi^* = (x_1^*, \ldots, x_n^*)$ e.g., $\xi^* = (0, \ldots, 0)$

$$f(x_Z) \triangleq \log P(x_Z, \xi^*_Z)$$

is defined for all $Z \subseteq \{1, \ldots, N\}$

$$f(x) = \log P(x)$$

Its Mobius inversion:

$$g(x_Z) = -\sum_{S \subseteq Z} (-1)^{|Z-S|} \log P(x_S, \xi^*_S)$$
Given a fixed assignment $\xi^* = (x_1^*, \ldots, x_n^*)$ e.g., $\xi^* = (0, \ldots, 0)$

$$f(x_Z) \triangleq \log P(x_Z, \xi_{-Z}^*)$$

is defined for all $Z \subseteq \{1, \ldots, N\}$

$$f(x) = \log P(x)$$

Its Mobius inversion: $g(x_Z) = -\sum_{S \subseteq Z} (-1)^{|Z-S|} \log P(x_S, \xi_{-S}^*)$

Define factors over each subset of nodes: $\psi_Z(x_Z) = -g(x_Z)$
Parametrization: **canonical form**

Given a *fixed assignment* $\xi^* = (x_1^*, \ldots, x_n^*)$ e.g., $\xi^* = (0, \ldots, 0)$

\[
f(x_Z) \triangleq \log P(x_Z, \xi_{-Z}^*)\] is defined for all $Z \subseteq \{1, \ldots, N\}$

\[
f(x) = \log P(x)\]

Its *Mobius inversion*: $g(x_Z) = -\sum_{S \subseteq Z}(-1)^{|Z-S|} \log P(x_S, \xi_{-S}^*)$

Define **factors** over each subset of nodes: $\psi_Z(x_Z) = -g(x_Z)$

From Mobius lemma: $P(x) = \exp\left(-\sum \psi_Z(x_Z)\right)$
Parametrization: **canonical form**

Its *Mobius inversion*: \( g(x_Z) = - \sum_{S \subseteq Z} (-1)^{|Z-S|} \log P(x_S, \xi^*_{S}) \)

Define **factors** over each subset of nodes: \( \psi_Z(x_Z) = -g(x_Z) \)

From Mobius lemma: \( P(x) = \exp(- \sum_{Z} \psi_Z(x_Z)) \)

**Problem:** one factor per subset of nodes

**Proof of Hammersly-Clifford theorem:**

When \( Z \) is not a clique \( \psi_Z(x_Z) \) becomes zero.
Proof of the **Hammersley-Clifford**

**Recap:**

- fix an assignment
- define factors over each subset of nodes as:

\[
\psi(x_Z) = \sum_{S \subseteq Z} (-1)^{|Z-S|} \log P(x_S, \xi_{-S})
\]

- if \( Z \) is not a clique in \( \mathcal{H} \) then \( \exists i, j \in Z \) such that \( X_i \perp X_j \mid X - \{X_i, X_j\} \)
  - we can show that \( \psi_Z(x_Z) = 0 \ \forall x_Z \)