

3 : Representation of Undirected GMs

Lecturer: Eric P. Xing

Scribes: Xing JunJie

1 Review

There are several important concepts and theorems introduced in last lecture about Directed Graphical Models.

- Local independence: For each variable $X_i : (X_i \perp \text{NonDescendant}_{X_i} | \text{Pa}_{X_i})$. Indicate that in a directed graph, each variable is independent to its nondescendants given its parent.
- Global independence: $I(G) = \{(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) : d\text{-sep}_G(\mathbf{X}; \mathbf{Y} | \mathbf{Z})\}$. The *global* independence is given by d-separation. Note that there is no need to consider too much about *global and local* things, you can call them whatever you want.
- A fully connected DAG \mathcal{G} is an I-map of *any* distribution, since $I_l(\mathcal{G}) = \emptyset \subset I(P)$ for any P .
- Minimal I-map: A DAG \mathcal{G} is a minimal I-map of P , if the removal of even a single edge from \mathcal{G} renders it not an I-map.
- A distribution may have several I-maps.
- P-map: A DAG \mathcal{G} is a perfect map (p-map) of a distribution P if $I(P) = I(\mathcal{G})$

Note that not every distribution has a perfect map as DAG. Here is an example:

$$A \perp C | \{B, D\} \quad B \perp D | \{A, C\}$$

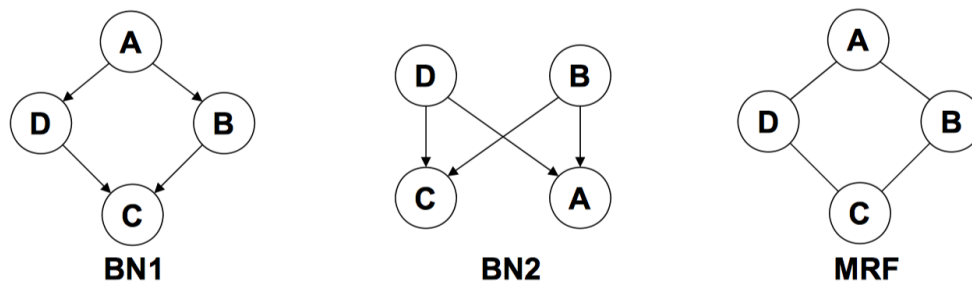


Figure 1: BN1 wrongly says $B \perp D | A$, BN2 wrongly says $B \perp D$

It is impossible for a DAG to capture both of the two independences at same time. The main reason is that the directed model (sometimes) encodes more independences together with the one we want. Thus, there is a portion of the space of distribution that we cannot encode with a DGM. That motivates another type of graphical model: undirected graphical models, aka Markov Random Fields.

2 Undirected Graphical Models

UGMs are very similar to DGMs in structure; but the directed or undirected edges encode differently. The directed model encodes *causal* relationship between nodes, while UGMs captures pairwise relationship which represents *correlation* between nodes, rough affinity.

Many things can be modeled as a UGM, such as a photo—each pixel can be a node, a go game—the grid chessboard seems intuitive, or even social networks, as shown in figure 2.

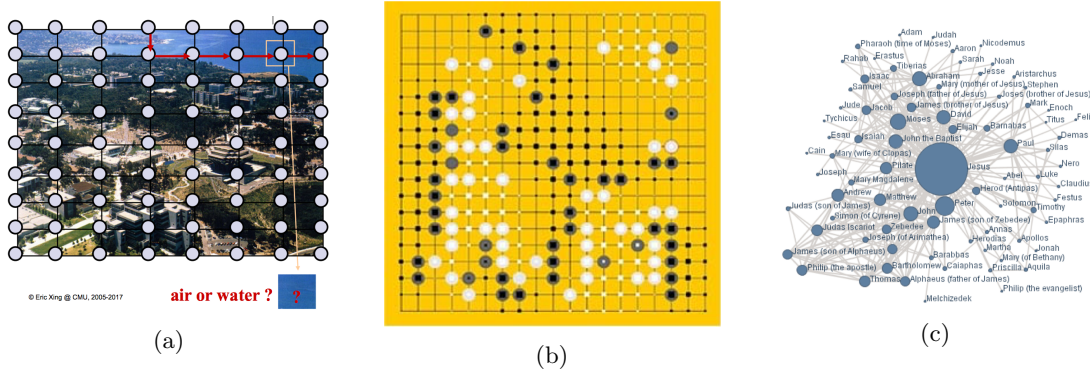


Figure 2: UGM examples

3 Representation

Definition 3.1 *an undirected graphical model represents a distribution $P(X_1, \dots, X_n)$ defined by an undirected graph H , and a set of positive potential functions ψ_c associated with the cliques of H , s.t.*

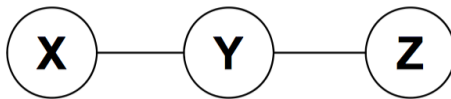
$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(X_c) \quad (1)$$

where Z is known as a partition function:

$$Z = \sum_{X_1, \dots, X_n} \prod_{c \in C} (\psi_c(X_c)) \quad (2)$$

The potential function can be understood as an contingency function of its arguments assigning “pre-probabilistic” score of their joint configuration. We call this of distribution in equation 1 as **Gibbs distribution**, as *Definition 4.3 in Koller textbook*. And the potential function is defined as **factor** in Koller textbook.

Definition 3.2 *For $G = V, E$, a complete subgraph (clique) is a subgraph $G' = V' \subseteq V, E' \subseteq E$ such that nodes in V' are fully interconnected. A (maximal) clique is a complete subgraph s.t. any superset $V'' \supset V'$ is not complete.*



3.1 Interpretation of Clique Potentials

The model implies $X \perp Z|Y$. This independence statement implies (by definition) that the joint must factorize as:

$$p(x, y, z) = p(y)p(x|y)p(z|y)$$

We can write this as

$$p(x, y, z) = p(x, y)p(z|y)$$

or

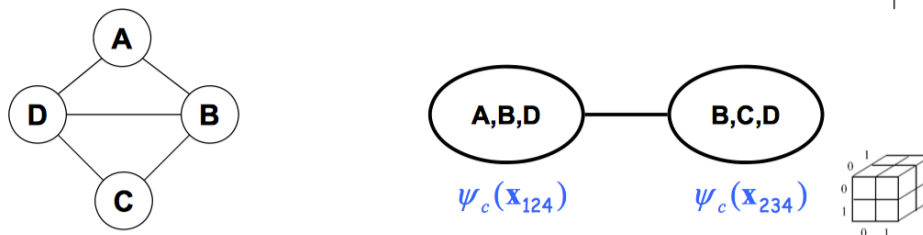
$$p(x, y, z) = p(x|y)p(z, y)$$

However, we cannot have all potentials be marginals and cannot have all potentials be conditionals.

The positive clique potentials can only be thought of as general “compatibility”, “goodness” or “happiness” functions over their variables, but not as probability distributions.

3.1.1 Example UGM — using max cliques

Here we’ll use an example to show an UGM.



We can factorize the graph into two max cliques:

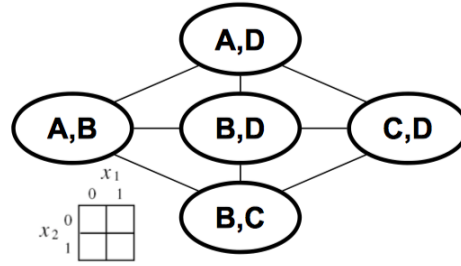
$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \psi_c(X_{123}) \times \psi_c(X_{234})$$

$$Z = \sum_{x_1, x_2, x_3, x_4} \psi_c(X_{123}) \times \psi_c(X_{234})$$

We can represent $P(X_{1:4})$ as two 3D tables instead of one 4D table.

3.1.2 Using subcliques

In this example, the distribution factorized over the subcliques.



$$\begin{aligned}
 P(x_1, x_2, x_3, x_4) &= \frac{1}{Z} \prod_{ij} \psi_{ij}(X_{ij}) \\
 &= \frac{1}{Z} \psi_{12}(X_{12}) \psi_{14}(X_{14}) \psi_{23}(X_{23}) \psi_{24}(X_{24}) \psi_{34}(X_{34}) \\
 Z &= \sum_{x_1, x_2, x_3, x_4} \prod_{ij} \psi_{ij}(X_{ij})
 \end{aligned}$$

3.1.3 Example UGM — canonical representation

A canonical representation of such a graph can be expressed as:

$$\begin{aligned}
 P(x_1, x_2, x_3, x_4) &= \frac{1}{Z} \psi_c(X_{123}) \times \psi_c(X_{234}) \\
 &\quad \times \frac{1}{Z} \psi_{12}(X_{12}) \psi_{14}(X_{14}) \psi_{23}(X_{23}) \psi_{24}(X_{24}) \psi_{34}(X_{34}) \\
 &\quad \times \psi_{x_1}(x_1) \psi_{x_2}(x_2) \psi_{x_3}(x_3) \psi_{x_4}(x_4) \\
 Z &= \sum_{x_1, x_2, x_3, x_4} \dots
 \end{aligned}$$

3.2 Independence properties

3.2.1 Global independence

Definition 3.3 A set of nodes Z separates X and Y in H , denoted $\text{sep}_H(X : Y|Z)$, if there is no active path between any node $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given \mathbf{Z} . Global independences associated with H are defined as:

$$I(H) = (X \perp Y|Z) : \text{sep}_H(X : Y|Z) \quad (3)$$

In Figure 3, B separates A and C if every path from a node in A to a node in C passes through a node in B. It is written as $\text{sep}_H(A : C|B)$. A probability distribution satisfies the global Markov property if for any disjoint A,B,C such that B separates A and C, A is independent of C given B.

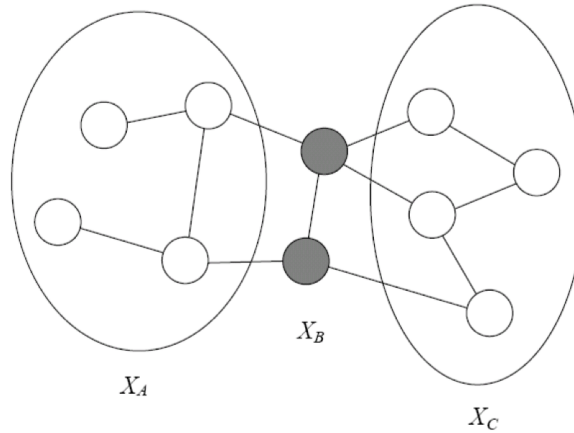


Figure 3: In this, the set X_B separates X_A from X_C . All paths from X_A to X_C pass through X_B

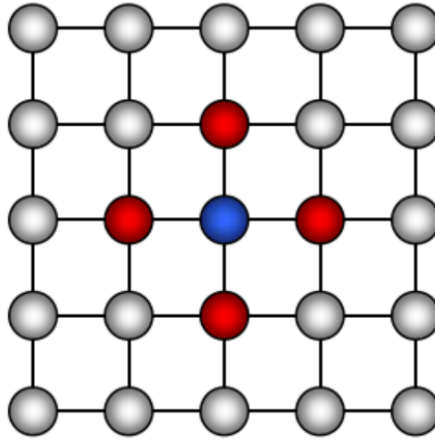


Figure 4: Illustration of Markov Blanket in undirected graph

3.2.2 Local independence

Definition 3.4 For each node $X_i \in V$, there is unique Markov blanket of X_i , denoted MB_{X_i} , which is the set of neighbors of X_i in the graph (those that share an edge with X_i)

Definition 3.5 The local Markov independencies associated with H is:

$$I_l(H) : \{X_i \perp V - \{X_i\} - MB_{x_i} | MB_{x_i} : \forall i\} \quad (4)$$

In other words, X_i is independent of the rest of the nodes in the graph given its immediate neighbors.

Note that, based on the local independence:

$$P(X_i | X_{-i}) = P(X_i | MB_{x_i}) \quad (5)$$



Figure 5: Pairwise independence in undirected graph. Red nodes are observed.

3.2.3 Soundness and completeness of global Markov property

Definition 3.6 An UG H is an I-map for a distribution P if $I(H) \subseteq I(P)$, i.e., P entails $I(H)$.

Definition 3.7 P is a Gibbs distribution over H if it can be represented as

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(X_c) \quad (6)$$

Theorem 3.1 (soundness): If P is a Gibbs distribution over H , then H is an I-map of P .

Theorem 3.2 (Completeness): If X and Y are not separated given Z in H ($\neg \text{sep}_H(X; Z|Y)$), then X and Y are dependent given Z , in some distribution P represented as $(X \perp_P Z|Y)$ that factorizes over H .

The proof of the theorems are available on Koller textbook.

3.2.4 Other Markov properties

For directed graphs, we defined I-maps in terms of local Markov properties, and derived global independence. For undirected graphs, we defined I-maps in terms of global Markov properties, and will now derive local independence.

The pairwise Markov independencies associated with UG $H = (V; E)$ are

$$I_p(H) = \{(X \perp Y | V - \{X, Y\}) : X, Y \notin E\}$$

For example, in figure 5, we have the following independence

$$X_1 \perp X_5 | \{X_2, X_3, X_4\}$$

3.2.5 Relationship between local and global Markov properties

- For any Markov Network H , and any distribution P , we have that if $P \models I_l(H)$ then $P \models I_p(H)$
- For any Markov Network H , and any distribution P , we have that if $P \models I_l(H)$ then $P \models I_p(H)$
- Let P be a positive distribution. If $P \models I_l(H)$, then $P \models I_p(H)$

The following three statements are equivalent for a positive distribution P :

- $P \models I_l(H)$

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

Above equivalence relies on the positivity assumption of P . For nonpositive distributions, there are examples of distributions P , there are examples which satisfies one of these properties, but not the stronger property.

3.2.6 Perfect maps

Definition 3.8 A Markov network H is a perfect map for P if for any $X; Y; Z$ we have that

$$sep_H(X; Z|Y) \Leftrightarrow P \models (X \perp Z|Y) \quad (7)$$

Note that, just like DMs, not every distribution has a perfect map as UGM.

3.2.7 Exponential Form

Constraining clique potentials to be positive could be inconvenient (e.g., the interactions between a pair of atoms can be either attractive or repulsive). We represent a clique potential $\psi_x(X_c)$ in an unconstrained form using a real-value "energy" function $\phi_c(X_c)$:

$$\psi_c(X_c) = exp\{-\phi_c(X_c)\} \quad (8)$$

Thus, this gives the joint distribution an additive structure:

$$P(X) = \frac{1}{Z} exp\left\{-\sum_{c \in C} \phi_c(X_c)\right\} = \frac{1}{Z} exp\{-H(X)\} \quad (9)$$

where the $H(X)$ is called the "free energy".

The exponential ensures that the distribution is positive. In physics, this is called the "Boltzmann distribution". In statistics, this is called a log-linear model (as Koller textbook introduces).