10-708: Probabilistic Graphical Models 10-708, Spring 2016

3 : Representation of Undirected GM

Lecturer: Eric P. Xing *Scribes: Longqi Cai, Man-Chia Chang*

1 MRF vs BN

There are two types of graphical models: one is Bayesian Network, which uses directed edges to model causality relationships with a Directed Acyclic Graph (DAG); the other is Markov Random Field (MRF), which uses undirected edges to model correlations between random variables. There are two difference between these two models.

Factorization rule: Bayesian Network (BN) uses chain rule, with each local conditional independence represented as factor:

$$\begin{aligned}&P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)\\ =\ &P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2)P(X_6|X_3, X_4)P(X_7|X_6)P(X_8|X_5, X_6)\end{aligned}$$

Markov Random Field (MRF) uses exponential of sums of energy functions, rather than simply production of factors.

$$\begin{aligned}&P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)\\ =\ &1/Z\ \exp\{E(X_1) + E(X_2) + E(X_3, X_1) + E(X_4, X_2)+\\ &E(X_5, X_2) + E(X_6, X_3, X_4) + E(X_7, X_6) + E(X_8, X_5, X_6)\}\end{aligned}$$

Partition Function: As each term of MRF does not have direct probabilistic meaning as BN, the sum of the exponential terms are not guaranteed to be one, so a partition function Z,which serves as a normalization factor, is needed to make this function a valid probability.

2 Independence in Graphical Model

2.1 I-map

Independence map (I-map) is defined as the independence properties encoded in the graph. Let $I(G)$ be the set of local independence properties encoded by a DAG G, then a DAG G is an I-map for distribution P if $I(G) \subseteq I(P)$. Thus, a fully connected DAG G is an I-map for any distribution since $I(G)$ is the empty set and is a subset of all possible $I(P)$.

A DAG G is a minimal I-map for distribution P if it is an I-map for P, and if a removal of even a single edge from G would make it not an I-map for P.

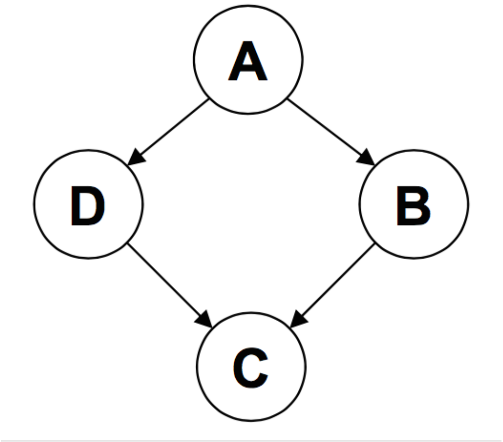

Figure 1: Bayes net 1.

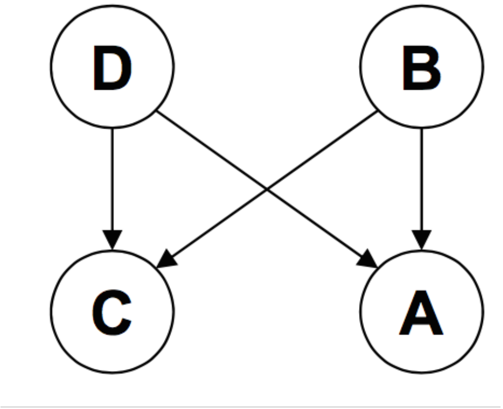

Figure 2: Bayes net 2.

2.2 P-map

A DAG G is a perfect map (P-map) for a distribution P if $I(G) = I(P)$. Not every distribution has a P-map as DAG. This can be proved by a counterexample.

Example: Suppose we have four random variable A, B, C, and D, and a set of conditional independence properties $I = \{A \perp C|\{B, D\}$ and $B \perp D|\{A, C\}\}$. These conditional independence properties can not be represented by any DAG. As shown in Figure 1, the Bayes net can represent $A \perp C|\{B, D\}$ and $B \perp D|A$ but can not represent $B \perp D|C$. As for Figure 2, the Bayes net can imply $A \perp C|\{B, D\}$, but can not imply $B \perp D|\{A, C\}$.

In the example, we can see none of the DAG can represent the conditional independencies, which proves that not every distribution has a perfect map as a DAG. Also, the fact that a graph G is a minimal I-map for a distribution P does not guarantee that G captures all the independence properties of P.

Nevertheless, we can find a MRF graph, which is shown in Figure 3, to represent all the conditional independencies in the example.

3 Applications of Undirected Graphical Model

Unlike a DAG, an undirected graphical model, also known as MRF, illustrates pairwise relationships rather than parental relationships. We can write down the model of undirected graphical model and score specific configuration of the graph, but there is no explicit way to generate sample from undirected graphical model.

Undirected graphical model is widely used in information retrieval and data mining realms. The grid model shown in Figure 4 is a canonical example of undirected graph. This model can be applied in image processing or lattice physics. Each node in the grid model could represent a single pixel or an atom. Due to continuity, adjacent or nearby nodes may have some relationship. For example, in Figure 5,it is really hard to say the small block, which at the up-right side of the image, is air or water. But if we look at the whole image, we can realize that the block is water. This is because we use the information from the nearby blocks.

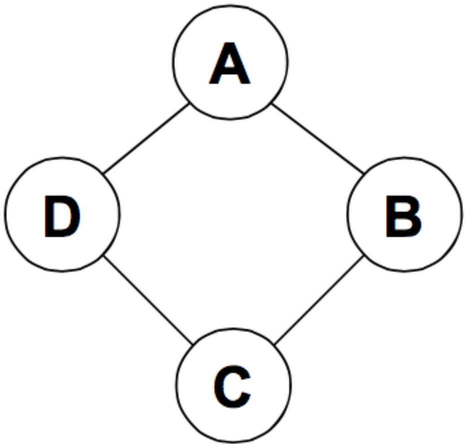

Figure 3: The MRF.

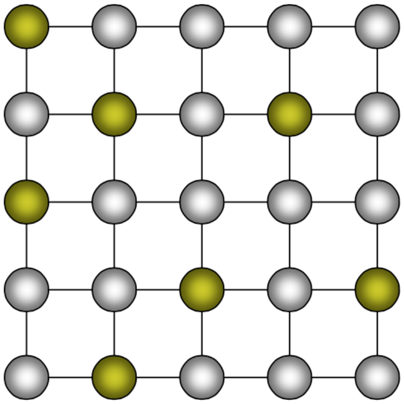

Figure 4: The grid model.

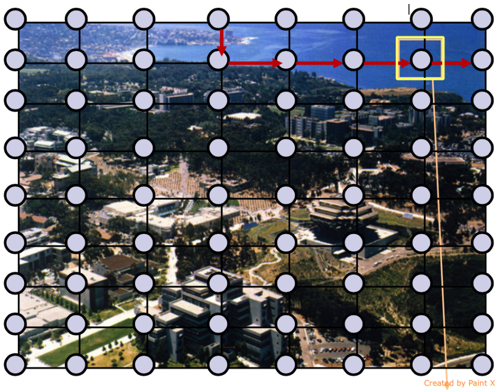

Figure 5: The Scene.

Besides, undirected graphical models can also be applied in social networks or protein interactive networks.

4 Markov Random Field (MRF)

- Gibbs distribution:

$$P(x_1, ..., x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$$

- Potential function: $\psi_c(x_c)$, where c corresponds to index of a clique.

- Partition function: $Z = \sum_{x_1,...,x_n} \prod_{c \in C} \psi_c(x_c)$

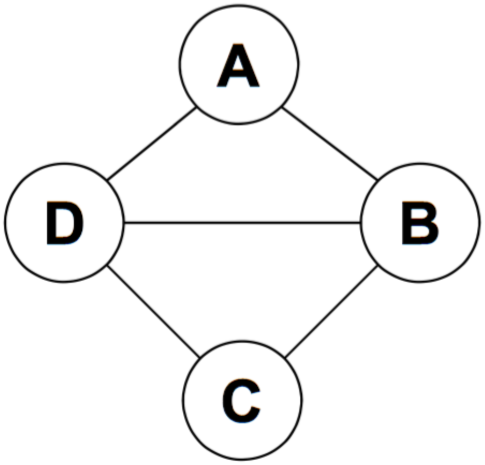

Figure 6: Clique demo

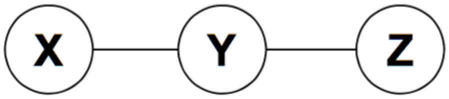

Figure 7: Counter example showing that potentials can be neither be viewed as marginals, nor as conditionals.

4.1 Clique

- **Clique**: for $G = \{V, E\}$, clique is a complete subgraph $G' = \{V' \subseteq V, E' \subseteq E\}$, such that nodes in V' are fully connected. For example in Figure 6, $\{A, B, D\}$ is a valid clique because $(A, B), (B, D), (D, A)$ connect all the nodes. However $\{A, B, C, D\}$ is invalid, because (A, C) is not connected.

- **Maximal clique**: a complete subgraph such that any superset $V'' \supseteq V'$ is not complete. For example $\{A, B, D\}$ is a maximal clique, because adding C in will cause the graph incomplete. $\{A, B\}$ is not maximal, because, adding D in, the graph will still be complete

- **Sub-clique**: Any subset of maximal clique will form a sub-clique. The minimal sub-cliques can be edges and singletons.

4.2 Clique potentials

- Potentials can be **neither** be viewed as marginals, **nor** as conditionals. Figure 7 is a counter example.

$$\begin{aligned} P(x, y, z) &= P(y)P(x|y)P(z|y) \\ &= P(x, y)P(z|y) \\ &= P(x|y)P(z, y) \end{aligned}$$

Given the conditional independences revealed by Figure 7, we can factorize the joint distribution as above. However, neither way can we factorize it into a form, such that each term corresponds to a potential function consistently. In this case, either one is joint, the other is conditional, or vice versa.

- Potentials should be thought as measure of ”compatibility”, ”goodness”, or ”happiness” over the assignment of a clique of variable. For example if $\psi(1, 1, 1) = 100, \psi(0, 0, 0) = 1$, it means the system is more compatible if $(X, Y, Z) = (1, 1, 1)$.

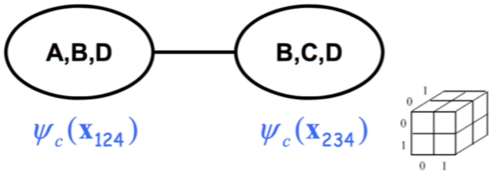

Figure 8: Max clique representation of Figure 6.

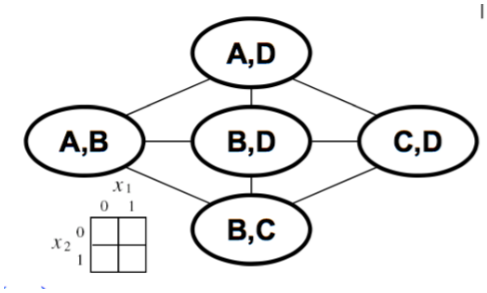

Figure 9: Sub clique representation of Figure 6.

4.3 Max clique vs sub-clique vs canonical representation

Note by: here x_1, x_2, x_3, x_4 are aliases for x_A, x_B, x_C, x_D.

- Canonical representation of Figure 6: factorize every possible clique.

$$\begin{aligned}P(x_1,x_2,x_3,x_4) = \frac{1}{Z}\,&\psi_{124}(x_{124})\psi_{234}(x_{234})\\ &\psi_{12}(x_{12})\,\psi_{14}(x_{14})\,\psi_{23}(x_{23})\,\psi_{24}(x_{24})\,\psi_{34}(x_{34})\\ &\psi_1(x_1)\,\psi_2(x_2)\,\psi_3(x_3)\,\psi_4(x_4)\end{aligned}$$

- Max clique representation (Figure 8): factorize over max cliques.

$$P'(x_1,x_2,x_3,x_4) = \frac{1}{Z}\,\psi_{124}(x_{124})\,\psi_{234}(x_{234})$$

- Sub clique representation (Figure 9): factorize over sub cliques.

$$P''(x_1,x_2,x_3,x_4) = \frac{1}{Z}\,\psi_{12}(x_{12})\,\psi_{14}(x_{14})\,\psi_{23}(x_{23})\,\psi_{24}(x_{24})\,\psi_{34}(x_{34})$$

- I-map: they represent the same graphical model, so the I-map should be the same.

$$I(P) = I(P') = I(P'')$$

- Distribution family: canonical form has the finest granularity, then comes sub-clique form, and then max-clique form, as can be seen from the number of parameters. Another interpretation is that sub-clique form can be marginalized into the max-clique form, but cannot be factorized back, so sub-clique encodes richer information than max clique. Similar analysis also applies to canonical form.

$$D(P) \supseteq D(P'') \supseteq D(P')$$

5 Independence in MRF

The independence properties of MRF can be implied from Markov property. There are two kinds of Markov property in MRF, global Markov property and local Markov property.

5.1 Global Markov Independence

In Figure 10, it can be seen that the set X_B separate all active paths from X_A to X_C, which can be denoted as $sep_H(A; C|B)$. This means every path from a node in set X_A to a node in X_C must pass through a node in X_B. A probability distribution satisfy global Markov independence if for any disjoint A, B, C, such that B separate A and C, then A is independent of C given B, which can be written as

$$I(H) = \{A \perp C|B : sep_H(A; C|B)\}$$

There are two theorems about global Markov independence. Let a MRF be H and a distribution P.

Soundness: If P is a Gibbs distribution over H,then H is an I-map for P.

Completeness: If $\neg sep_H(X; Z|Y)$, there are some P that factorizes over H such that $X \not\perp_P Z|Y$.

5.2 Local Markov Independence

There is unique Markov blanket of each node in a MRF. Take Figure 11 for example. The Markov blanket of $X_i \in V$,denoted MB_{X_i},is the set of direct neighbors of X_i which share edge with X_i. The local Markov independence of the graph in Figure 11 is defined as:

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

That is, X_i is independent of all other nodes in the graph given its direct neighbors.

5.3 Relation Between Global and Local Markov Independence

For MRF, we can also define local pair-wise Markov independencies associated with H as follow:

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

For example, a pairwise Markov independence in Figure 12 can be described as $X_1 \perp X_5|\{X_2, X_3, X_4\}$.

There are several relationships between global and local Markov properties.

- If $P \models I_l(H)$, then $P \models I_p(H)$.
- If $P = I(H)$, then $P \models I_l(H)$.
- If $P > 0$ and $P \models I_p(H)$, then $P \models I_l(H)$.
- The following statements are equivalent for a *positive* distribution P:

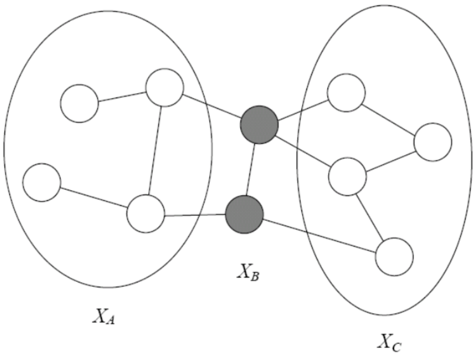

Figure 10: The global Markov properties.

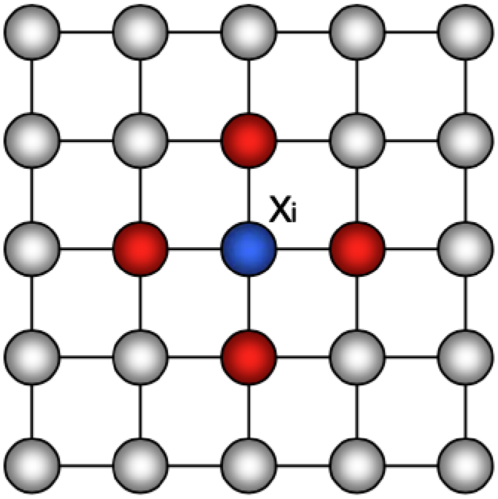

Figure 11: The local Markov properties.

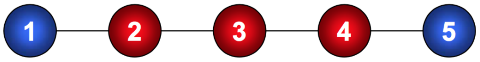

Figure 12: The pairwise Markov properties.

$$\begin{gathered} P \models I_l(H) \\ P \models I_p(H) \\ P \models I(H) \end{gathered}$$

5.4 Hammersley-Clifford Theorem

We have so far represented undirected graphical model by Gibbs distribution, which is a product factorization of potential functions. The conditional independence properties of undirected graphical model can be described in Markov properties. The Hammersley-Clifford theorem states that a positive probability distribution can be factorized over the cliques of the undirected graph. That is, all joint probability distribution that satisfies the Markov independence can be written as potentials over maximal cliques, and given a Gibbs distribution, all of its joint distributions satisfy the Markov independence. The formal theorem is illustrated following:

Theorem: Let P be a *positive* distribution over V, and H a Markov network graph over V, If H is an I-map for P, then P is a Gibbs distribution over H.

5.5 Perfect Map

A Markov network H is a perfect map for a distribution P if for any X, Y, Z, we have that $sep_H(X; Z|Y) \Leftrightarrow P \models (X \perp Z|Y)$. Not every distribution has a perfect map as undirected graphical model, just the same as

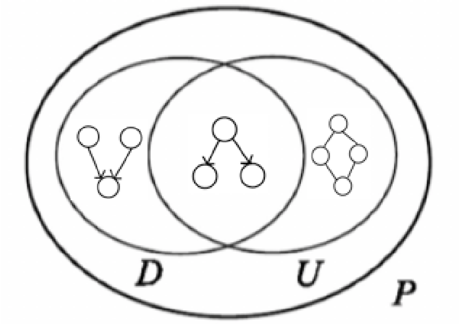

Figure 13: The distribution coverage of Bayesian network and MRF.

what we described earlier in Section 2. A counterexample can be given. The independencies encoded in a v-structure $X \rightarrow Z \leftarrow Y$ can not be captured by any undirected graphical model.

There are some distributions that can be captured by DAG while some can be captured by MRF. Figure 13 is a venn diagram which illustrates the distribution that DAG and MRF can capture, provided with example graphs.

5.6 Exponential Form

Constraining clique potentials to be positive could be inconvenient,for example the interactions between atoms can be attractive or repulsive. To address this problem, we can represent a clique potential $\psi_c(x_c)$ in an unconstrained form using a real-value energy function $\phi_c(x_c)$ with exponential forms

$$\psi_c(x_c) = \exp\{-\phi_c(x_c)\}$$

The exponential forms provides a nice additive property in that we can write the distribution as:

$$p(x) = \frac{1}{Z} \exp\Big\{ - \sum_{c \in C} \phi_c(x_c) \Big\} = \frac{1}{Z} \exp\Big\{ - H(x) \Big\}$$

where the sum in the exponent is called the "free energy":

$$H(x) = \sum_{c \in C} \phi_c(x_c)$$

The form of distribution $p(x)$ has different names in different realms. In physics, this is called the "Boltzmann distribution", while in statistics, this is called a log-linear model.

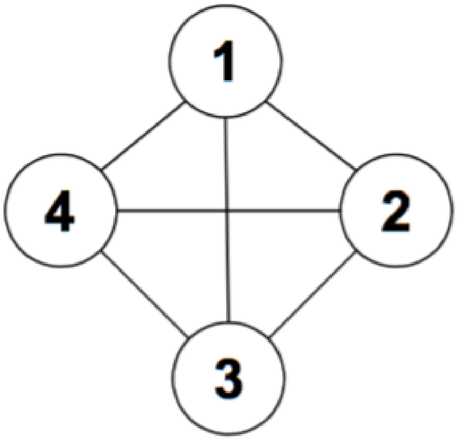

Figure 14: Bolzman Machine example

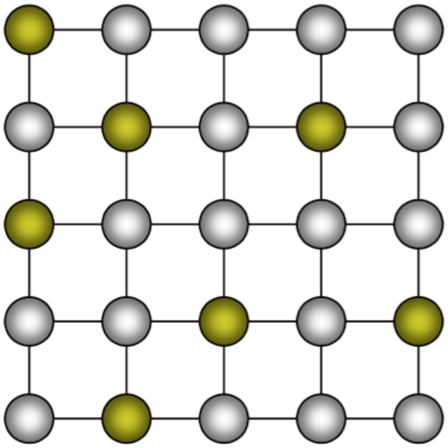

Figure 15: Ising model

6 Examples

6.1 Boltzman Machine

Definition: fully connected graph with pairwise potentials on binary-valued nodes ($x_i \in \{-1, +1\}$ or $\{0, 1\}$)

$$\begin{aligned} P(x_1, x_2, x_3, x_4) &= \frac{1}{Z} \exp\Big\{ \sum_{ij} \phi_{ij}(x_i, x_j) \Big\} \\ &= \frac{1}{Z} \exp\Big\{ \sum_{ij} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i + C \Big\} \end{aligned}$$

Energy function in matrix form:

$$\begin{aligned} H(x) &= \sum_{ij} (x_i - \mu) \Theta_{ij} (x_j - \mu) \\ &= (x - \mu)^T \Theta (x - \mu) \end{aligned}$$

6.2 Ising model

Description: Nodes are arranged in grid and connected only to geometric neighbors (Figure 15).

$$P(X) = \frac{1}{Z} \exp\Big\{ \sum_{i,j \in N_i} \theta_{ij} X_i X_j + \sum_i \theta_{i0} X_i \Big\}$$

Potts model: when each node has multiple states.

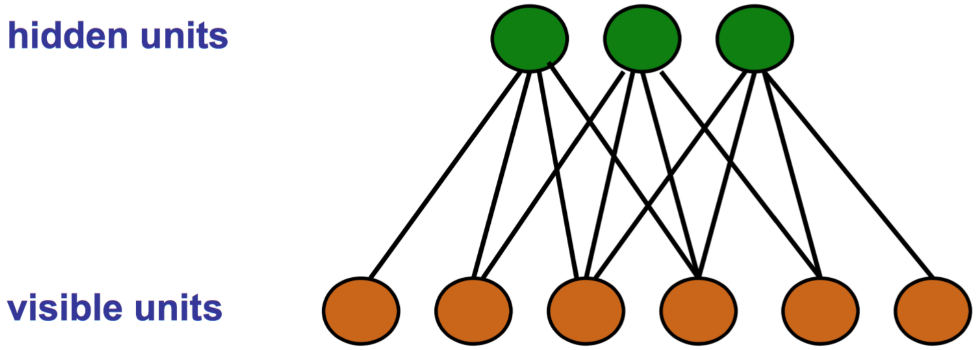

Figure 16: Restricted Bolzman Machine

6.3 Restricted Bolzman Machine (RBM)

Restricted in the way that it's a bipartite graph from hidden units to visible units.

$$p(x, h \mid \theta) = \exp\Big\{\sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{i,j} \phi_{i,j}(x_i, h_j) - A(\theta)\Big\}$$

6.3.1 Properties of RBM

- Factors are **marginally dependent**.
- Factors are **conditionally independent** given visible nodes.

$$P(h|x) = \prod_i P(h_i \mid x)$$

- Property of conditional independence makes it possible for iterative Gibbs sampling.

6.4 Conditional Random Field (CRF)

Figure 17 shows a transition from Hidden Markov Model (HMM) to CRF. First change the directed edges into undirected ones. Second, as we don't assume indepedences among features, we may merge all the features together, and labeling feature X_i, we would take all the features into account. Unlike HMM, CRF is a discriminative model, as it models posterior directly.

$$P_\theta(y|x) = \frac{1}{Z(\theta, x)} \exp\Big\{\sum_c \theta_c f_c(x, y_c)\Big\}$$

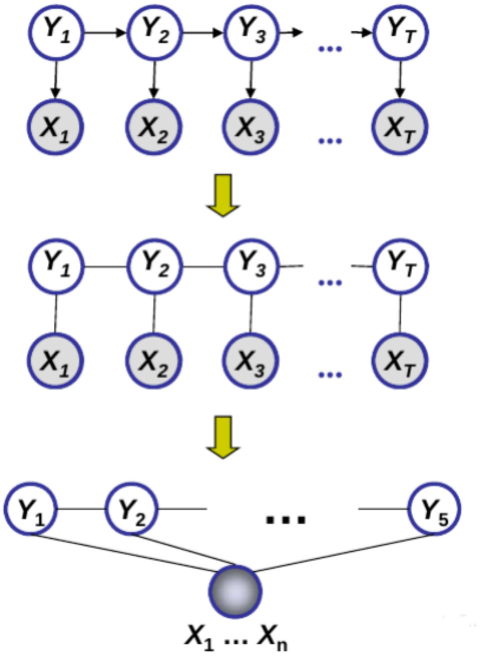

Figure 17: Conditional Random Field

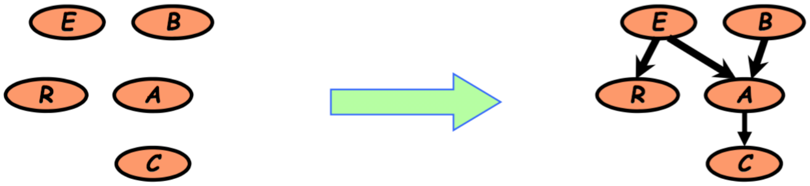

Figure 18: Structure learning

7 Structure Learning

The problem of structure learning is that, given a set of independent samples, find the best graphical model topology. The goal is to optimize the following likelihood function.

$$\begin{aligned}\ell(\theta_G, G; D) &= \log \hat{p}(D \,|\, \theta_G, G) \\ &= M \sum_i \hat{I}(x_i, \mathbf{x}_{\pi_i(G)}) - M \sum_i \hat{H}(x_i)\end{aligned}$$

The problem is difficult because there are $O(2^{n^2})$ graphs and $O(n!)$ trees over n nodes. However, we are able to find the exact solution of an optimal tree under MLE. It's solved by Chow-Liu algorithm. The trick is based on the fact that each node has only one parent in a tree.