6.438: Recitation-2 (9/22/2014)

I. Graphical Model Definitions

2 ways of defining graphical models:

1. Independence
2. Factorization

Directed Graphical Models:

Factorization: $P(X_1, X_2, \ldots X_N) = \prod_{i=1}^{N} P(X_i | X_{\Pi_i})$

Global Independences: $X_A \perp X_B \mid X_C$ if C *D-separates* A and B in G.

For directed models, the 2 definitions are equivalent.

Undirected Graphical Models:

In terms of Independences:

Global Independences: $X_A \perp X_B \mid X_C$ if removing C disconnects A from B in G.

Local Independences: $X_i \perp X_{V-\{i\}-N(i)} \mid X_{N(\mathrm{i})}$, where N(i) represents neighbors of i in G.

Pairwise Independences: $X_i \perp X_j \mid X_{V-\{i\}-\{j\}}$ if i and j do not have an edge in G.

In terms of Factorization: Over Maximal Cliques

$P(X_1, \ldots X_N) \approx \prod_{C=1}^{\mathcal{C}} \phi_C(X_C)$

$P(X_1, \ldots X_N) = \frac{1}{Z} * \prod_{C=1}^{\mathcal{C}} \phi_C(X_C)$

Where Z is a normalization factor called the Partition Function.

The 2 definitions are not equivalent; we need additional assumptions to get factorization from Independences, such as having a positive distribution (Hammersely-Clifford). The primary definition of undirected models is thus in terms of global CIs*.

*CI stands for Conditional Independence, I often use this interchangeably with Independence

The 3 types of independencies are also not equivalent. In general, Global Independences are stronger than Local Independences, which are in turn stronger than Pairwise Independences. i.e.
Global => Local => Pairwise ..(1)
(The first implication is actually a set containment relation, the second follows from basic properties of independence)

Moreover, in the absence of any other assumptions, these relations are strict i.e. there exist distributions which satisfy pairwise CIs of a graph but not local CIs (for some graphs), and similarly for local vs. global. (Remember, you have shown this in problem 4 in Pset-2!). However, if we are given the additional assumption that the distribution is positive, these 3 sets of independencies become equivalent. The easiest way to prove this is via the Hammersley-Clifford theorem:

In the Hammersley-Clifford theorem, we only make use of pairwise independencies to prove the existence of a factorization. (I would strongly encourage you to look at the proof and verify this). Thus, for a positive distribution, we have:
Pairwise Independence => **Factorization over Maximal Cliques ..(2)**

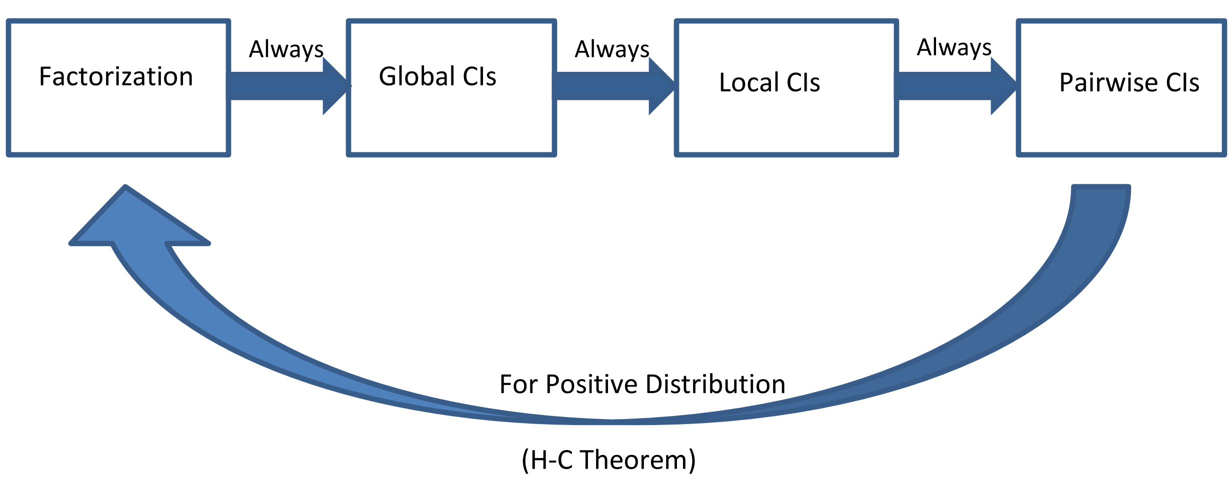
It can be easily shown that factorization is a stronger property than global independencies, i.e.
Factorization over Maximal Cliques => **Global Independences ..(3)**

(This is true for an arbitrary distribution i.e. we do not need to assume positivity. You may try to find a proof of the above statement as an exercise.)

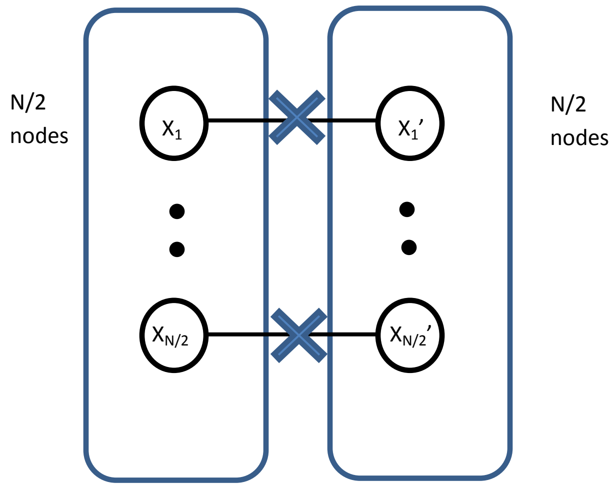
Combining (1), (2) and (3); we find that for a positive distribution, all the 4 forms of representation – factorization, global CIs, local CIs, and pairwise CIs – are equivalent.

In this course, we will mostly be dealing with a distribution that factorizes over the graph, so we will frequently assume positivity in lectures. In practice, any distribution can be converted to a positive distribution which is very close to the original, so we don't lose anything by this assumption.

Summary:



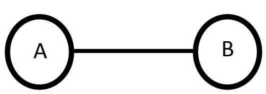
II. Maximal Cliques example



The above graph has all possible edges **except** the ones shown in the diagram.
For a graph G with N nodes, this method produces 2^(N/2) maximal cliques. Detailed explanation is left to the students.

III. Undirected graphical models examples (Not discussed in recitation)

Example 1:



We have 2 Binary random variables, X_A and X_B.

Node Potentials: $\phi_A(X_A) = 1,\ \phi_B(X_B) = 1$

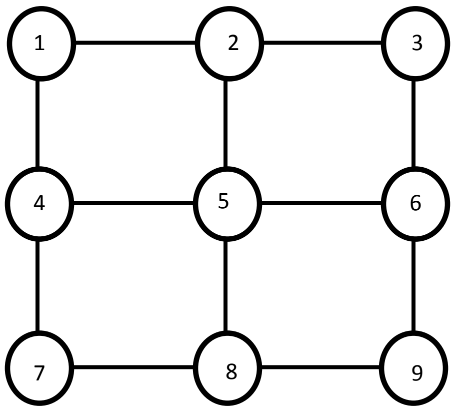
Edge Potentials: $\phi_{A,B}(x_A, x_B) = \begin{cases} 10, & x_A = \ x_B \\ 5, & x_A = 1 \ and \ x_B = 0 \\ 15, & x_A = 0 \ and \ x_B = 1 \end{cases}$

Thus, $P_{X_A,X_B}(x_A, x_B) = \frac{1}{Z}\ \phi_A(x_A)\,.\,\phi_B(x_B).\,\phi_{A,B}(x_A, x_B)$

Calculate:

i) Z
ii) Pr. (X_A = 1,X_B = 0) = ?
iii) PR. (X_B = 1 | X_A = 0) = ?

Example 2: Undirected graph for noise removal



The above graph represents a portion of an image, where each pixel takes on a binary value (0/1). Just as in a real image, adjacent pixels have high probability of having the same value. We capture this tendency using *edge potential functions*. In addition to this, we have certain "measurements" for the pixels which also influence our belief about their true values. These measurements are represented using *node potentials*.

To keep the setting simple, we assume that neighbors of 5 are observed exactly i.e. their node potentials are deterministic functions. Using this information, we try to determine the value of pixel 5.

Edge Potentials: $\forall (i,j) \in E$: $\phi_{i,j}(x_i, x_j) = \begin{cases} 1, & if\ x_i = x_j \\ 0.2, & otherwise \end{cases}$

Node Potentials: $\phi_5(x^5) = \begin{cases} 0.9, & if\ x^5 = 1 \\ 0.1, & if\ x^5 = 0 \end{cases}$

$\forall i \in V \backslash \{5\}$: $\phi_i(x_i) = \mathbb{I}(x_i = 0)$

Questions:
i) What is Pr.(X_5 = 1) ?

ii) Instead of knowing the neighboring nodes exactly, suppose that node potentials for these nodes are also of the same form as x_5, but inverted i.e. $\phi(0) = 0.9$, $\phi(1) = 0.1$. Can you now calculate P(x_5 = 1) exactly? If you had to fix a single value for x_5, either 0 or 1, which would you choose? (answer intuitively)

iii) (Challenge) What is the complexity of calculating Pr (X_5) exactly?
There is a naïve way to do it, which involves summing over all the variables. Can you do something better? What is the lowest complexity you can get for this operation? (Note: complexity here ≈ no. of operations).

Comments: Problem (ii) is a problem of finding the MAP assignment, and problem (iii) is finding the marginal probability (or inference problem). We will see that (iii) is harder than (ii) in a formal sense.

IV: Factor graphs vs. Undirected graphical models

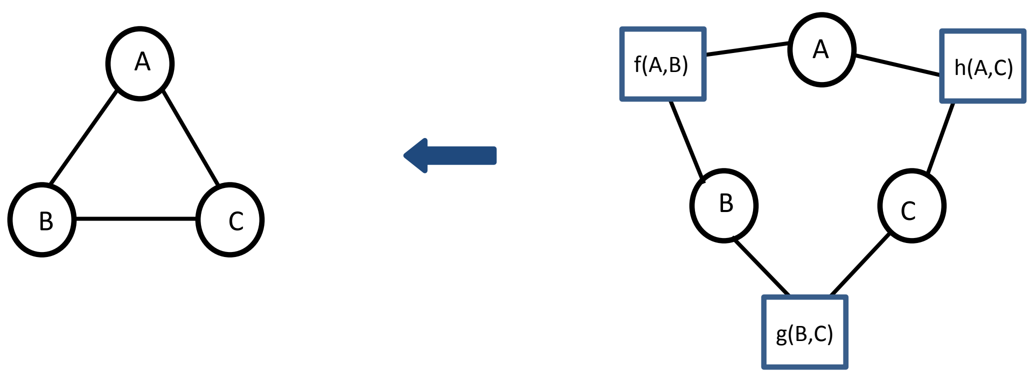
Undirected → Factor Graph:

Every undirected graph can be represented by an equivalent Factor graph. The way to do this is to create a factor graph with same set of variable nodes, and one factor node for each maximal clique in the graph. (We have assumed factorization/positivity of the distribution here.)

Factor Graph → Undirected:

The transformation from factor graph to undirected graph is not lossless in general. We get the undirected graphical model by reversing the above process - for each factor node, connect all variable nodes into a single clique.

We give an example below where such a transformation loses information.



G: 3-node undirected graphical model over A,B, C

G': Factor Graph over A, B, C (with edge potentials only)

On the right, we have a factor graph G', which leads to a factorization over edges. However, the graph on the left simply gives us a factor $\phi_{A,B,C}(X_A, X_B, X_C)$ which represents an arbitrary probability distribution over X_A, X_B, X_C.

The fact that these 2 are not equivalent can be shown by considering the following distribution, which can be represented by the graphical model on the left but not the one on the right:

$P_{X_A,X_B,X_C}(1,0,1) = \frac{1}{3}$; $P_{X_A,X_B,X_C}(1,1,0) = \frac{1}{3}$; $P_{X_A,X_B,X_C}(0,1,1) = \frac{1}{3}$

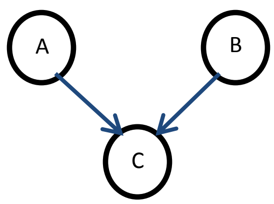
$P_{X_A,X_B,X_C}(1,1,1) = 0$

V. Directed graphical models vs. undirected graphical models

Question: Can we represent any directed graphical model using an undirected one, and vice versa?

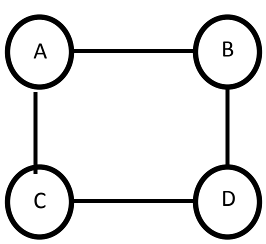
Answer: No. Directed and undirected graphical models are "different". There are certain distributions that have a directed P-map but no undirected P-map, and vice versa.

Example:
i) Directed Graph with V-structure.



This has no undirected graphical model as its P-map.

ii) Undirected graph with induced 4-cycle or longer cycle. (induced cycle is a chordless cycle)



This has no directed graphical model as its P-map.

VI. Discussion about P-maps, I-maps, D-maps

This is already covered in lecture-notes (Lecture-4).

6.438 Algorithms for Inference
Fall 2014