CS726 Scribe Notes

Saksham (22B1003), Sharvanee (22B0943), Deeksha (22B0988)
Department of Computer Science,
Indian Institute of Technology Bombay

1 Class Query

Based on one of the questions raised in class, ma'am clarified that:

$$X \perp\!\!\!\perp \{Y_1, \ldots, Y_k\} \mid Z \implies (X \perp\!\!\!\perp Y_1|Z) \ldots (X \perp\!\!\!\perp Y_k|Z)$$

However, the reverse might not be true. So, the following does not hold:

$$(X \perp\!\!\!\perp Y_1|Z) \ldots (X \perp\!\!\!\perp Y_k|Z) \implies X \perp\!\!\!\perp \{Y_1, \ldots, Y_k\} \mid Z$$

2 Undirected Graphical Models

In this model, the underlying graph which represents the probability distribution is undirected. Such a model is useful when the variables interact symmetrically, and there is no natural parent-child relationship. Some of the natural examples, which represent such a scenario include friends on a social media platform, atoms in a crystal, labelling pixels in an image, etc.

In an undirected graphical model, we define potentials over arbitrary cliques of the graph G. (Cliques are subsets of nodes of a graph which are fully connected, i.e they are complete subgraphs of a graph.) The potentials are denoted by $\psi_C(y_C)$. The first subscript C denotes the clique in consideration, and y_C denotes the assignment to the variables in the clique.

Potentials can take arbitrary non-negative values, however they cannot be considered equivalent to probabilities.

Here is how we define the joint distribution in an undirected graphical model:

$$Pr(y_1, \ldots, y_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(y_C)$$

where $\mathcal{C}$ is the set of all cliques in the graph, and Z is the normalizing constant, which ensures that the sum of all probabilities is equal to 1. For the numerator, what we essentially do is for that particular assignment of variables, we take a product over all the cliques in the graph, and multiply the potentials for that assignment.

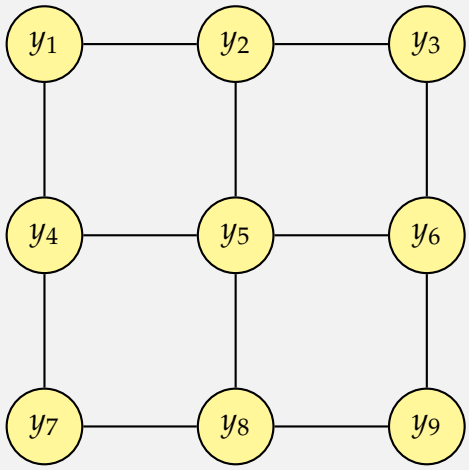
Here is how we define Z:

$$Z = \sum_{y_1} \cdots \sum_{y_n} \prod_{C \in \mathcal{C}} \psi_C(y_C)$$

This expression of Z is also called the partition function. For calculating Z, what we essentially do is to sum over all possible assignments of variables, and take a product over all the cliques in the graph, and multiply the potentials for that assignment. This ensures that the sum of all probabilities is equal to 1.

2.1 Example

Consider the graph G shown on the right. There are 9 binary variables $y_1, y_2, \ldots, y_9$ (each can take only two values 0 or 1).
There are two types of cliques in this graph, one is the set of all edges, and the other is the set of all nodes. The potentials for the edges are denoted by $\psi_{ij}(y_i, y_j)$, and the potentials for the nodes are denoted by $\psi_i(y_i)$.
The joint distribution for this graph can be written as:



$$Pr(y_1, \ldots, y_9) = \frac{1}{Z} \prod_{i=1}^{9} \psi_i(y_i) \prod_{(i,j) \in E} \psi_{ij}(y_i, y_j)$$

where E is the set of all edges in the graph.

For example, the value of $Pr(1, 0, \ldots, 0)$ can be calculated as:

$$\begin{aligned} Pr(1,0,\ldots,0) = \frac{1}{Z} & \psi_1(1)\psi_2(0)\psi_3(0)\ldots\psi_9(0)\psi_{12}(1,0)\psi_{23}(0,0)\psi_{14}(1,0)\psi_{25}(0,0) \\ & \psi_{36}(0,0)\psi_{45}(0,0)\psi_{56}(0,0)\psi_{47}(0,0)\psi_{58}(0,0) \\ & \psi_{69}(0,0)\psi_{78}(0,0)\psi_{89}(0,0) \end{aligned}$$

Z can be calculated as follows:

$$\begin{aligned} Z = \sum_{y_1} \cdots \sum_{y_9} & \psi_1(y_1)\psi_2(y_2)\psi_3(y_3)\ldots\psi_9(y_9)\psi_{12}(y_1,y_2)\psi_{23}(y_2,y_3)\psi_{14}(y_1,y_4)\psi_{25}(y_2,y_5) \\ & \psi_{36}(y_3,y_6)\psi_{45}(y_4,y_5)\psi_{56}(y_5,y_6)\psi_{47}(y_4,y_7)\psi_{58}(y_5,y_8) \\ & \psi_{69}(y_6,y_9)\psi_{78}(y_7,y_8)\psi_{89}(y_8,y_9) \end{aligned}$$

In this case, we need to sum over 512 possible assignments of variables, which is computationally expensive.

Another example was given. It was a K3 (complete graph with 3 nodes). If we consider only one clique (the complete graph), then the joint distribution can be written as (an example):

$$P(y_1 = 0, y_2 = 1, y_3 = 1) = \frac{\psi_{123}(0,1,1)}{\psi_{123}(0,0,0) + \psi_{123}(0,0,1) + \cdots + \psi_{123}(1,1,1)}$$

Similarly, if we consider only the three edges as the cliques, then the joint distribution can be written as:

$$P(y_1 = 0, y_2 = 1, y_3 = 1) = \frac{\psi_{12}(0,1)\psi_{23}(1,1)\psi_{13}(0,1)}{\psi_{12}(0,0)\psi_{23}(0,0)\psi_{13}(0,0) + \cdots + \psi_{12}(1,1)\psi_{23}(1,1)\psi_{13}(1,1)}$$

The types of cliques, on which we define potential depends on the real world clue which we have.

Also, the number of parameters needed to define potential is exponential in the number of nodes in the clique. Therefore, computing Z is tough and for certain graphs we exploit factorization to simplify.

In general, if $|C| = k$, and $y_i \in \{1, \ldots, m\}$, then the number of potential scores we need to report will be m^k.

3 Conditional Independencies in an Undirected Graphical Model

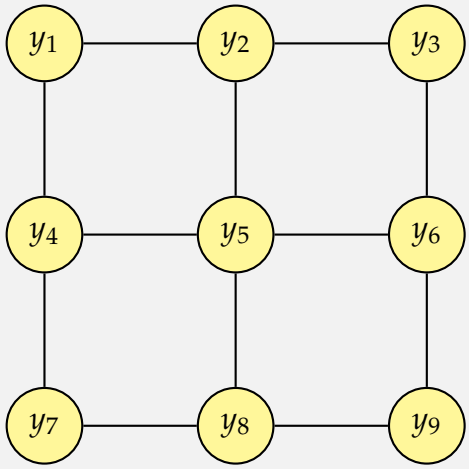
Let $V = \{y_1, \ldots, y_n\}$.

Let distribution P be represented by an undirected graphical model G. If Z separates X and Y in G, then $X \perp\!\!\!\perp Y|Z$ in P. The set of all such CIs are called Global-CI of the UGM (undirected graphical model).

Example:

1. $y_1 \perp\!\!\!\perp y_3, y_5, y_6, y_7, y_8, y_9 | y_2, y_4$
2. $y_1 \perp\!\!\!\perp y_3 | y_2, y_4, y_5, y_6, y_7, y_8, y_9$ (the separator Z need not be minimal)
3. $y_1, y_2, y_3 \perp\!\!\!\perp y_7, y_8, y_9 | y_4, y_5, y_6$
4. $y_1 \perp\!\!\!\perp y_3 | y_2, y_4$

(Basically when we remove Z, X and Y should be disconnected in the graph.)



4 Factorization implies Global CI

Here is the theorem:

Let G be the undirected graph over $V = x_1, \ldots, x_n$ nodes and $P(x_1, \ldots, x_n)$ be a distribution. If P is represented by G that is, if it can be factorized as per the cliques of G, then P will also satisfy the global-CIs of G.

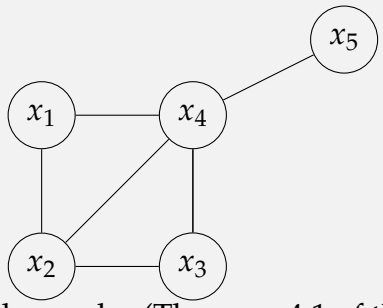
Factorize$(P, G) \implies$ Global-CI(P, G)

4.1 Example

One of the valid options to express P is as follows:

$$P(x_1, \ldots, x_5) \propto \psi_{124}(x_1, x_2, x_4)\psi_{234}(x_2, x_3, x_4)\psi_{45}(x_4, x_5)$$

Since, P can be factorized as per the cliques of G, it will also satisfy the global-CIs of G. For example, $x_1 \perp\!\!\!\perp x_5 | x_2, x_3, x_4$ is a valid CI.



The proof of this theorem has been left asn exercise for the reader (Theorem 4.1 of the KF book).

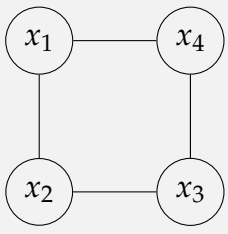
4.2 Global CI does not imply Factorization

The counter example wasn't discussed in class, but was left for the students to read through the slides. It has been taken from example 4.4 of the KF book.

Consider a distribution over 4 binary variables $P(x_1, x_2, x_3, x_4)$. The graph G is shown on the right.

Let $P(x_1, x_2, x_3, x_4)$ be $\frac{1}{8}$ when x_1, x_2, x_3, x_4 takes values from this set: $\{0000, 1000, 1100, 1110, 1111, 0111, 0011, 0001\}$. In all other cases it is 0. It is left as an exercise for the reader to check that all four global CIs hold in the graph: $x_1 \perp\!\!\!\perp x_3 | x_2, x_4$ etc.

Now, we will look at factorization. The factors corresponding to the edges in $\psi(x_1, x_2)$. Each of the four possible assignments of each factor is non-zero (as per the set of values mentioned above). But, this cannot represent the zero probability for cases like $x_1, x_2, x_3, x_4 = 0101$.

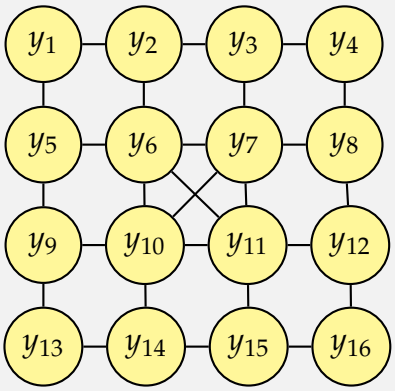


5 Drawing an undirected graphical model

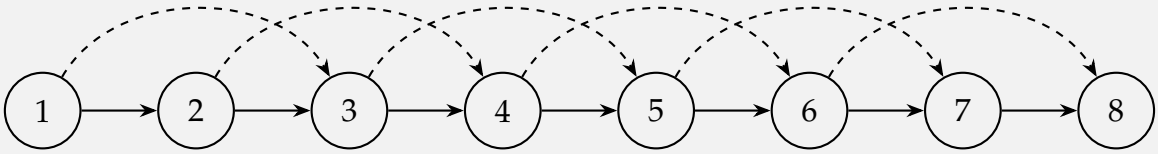
There are majorly two ways to draw an undirected graphical model:

1. **Starting from factors:** We simply connect together all variables that appear together in a factor. Here are some real-life examples:

   - Image pixels



   - Language Models from n-gram scores



   The figure above illustrates a 3-gram model, where nodes separated by a distance of 2 are connected. This ensures that contiguous groups of three words are linked, allowing the model to capture the context and meaning within each phrase.

2. **Starting from CIs**

6 Constructing an UGM from a positive distribution

Positive distribution is a distribution where all the probabilities are non-negative. We are given $P(x_1, \ldots, x_n)$ to which we can ask any CI of the form $X \perp\!\!\!\perp Y|Z$ and get a yes, no answer.

Our goal is to draw a minimal, correct UGM G to represent P. Here are the two options which we have (V denotes the set of all n variables):

- **Using pairwise CI:** For each pair of vertices x_i, x_j, if $x_i \not\perp\!\!\!\perp x_j|V - \{x_i, x_j\}$ in P, we add an edge between x_i and x_j in G. This is because for a UGM, the following is true:

$$X \perp\!\!\!\perp Y|Z \implies X \perp\!\!\!\perp Y|Z, W$$

  The above might not be true for a bayesian network. It is true for a UGM because even after adding nodes to Z, X and Y will still be disconnected in the graph. Hence considering the entire set $V - \{x_i, x_j\}$ will also work.

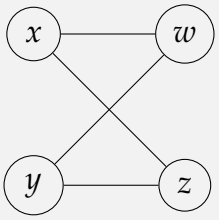
- **Using local CI:** For each node x_i, we need to find the smallest subset U such that $x_i \perp\!\!\!\perp V - U - \{x_i\}|U$ in P. After this, we make the nodes in U, the neighbours of x_i in G.

6.1 Example

We are given the positive distribution $P(x, y, z, w)$ for which the following CIs hold:

- $x \perp\!\!\!\perp y | z, w$
- $z \perp\!\!\!\perp w | x, y$

Let us first apply pairwise-CI algorithm to draw the graph G. We need to iterate over all the pairs of variables and check if the CI holds. If it does not hold, we add an edge between the two variables. So, for the edge x, y, they are independent given the other two variables ($x \perp\!\!\!\perp y | z, w$), so we do not add an edge. Similarly, for the edge z, w, they are independent given the other two variables ($z \perp\!\!\!\perp w | x, y$), so we do not add an edge. For all other edges, we add an edge (because the CI does not hold). So, the graph G will look like this:



Now, let us try the same problem with the local-CI algorithm. We need to find U for each of the four nodes. For x, U is $\{z, w\}$ because $x \perp\!\!\!\perp V - U - \{x\} | U$ holds. For y, U is $\{z, w\}$, for z, U is $\{x, y\}$, for w, U is $\{x, y\}$. We need to connect each node to the nodes in U. So, we will get the same graph as above.

6.2 Markov Blanket

UGMs are also called Markov Random Fields.

The Markov Blanket of a variable x_i, $MB(x_i)$ is the smallest subset of variables V that makes x_i, CI of others given the Markov Blanket. This is essentially the U, which we used above.

$$x_i \perp\!\!\!\perp V - MB(x_i) - \{x_i\} | MB(x_i)$$

Also, one of the theorems says that $MB(x_i)$ is always unique for a positive distribution. The proof of this, has been left as a self-reading exercise (given in the slides).

6.3 Hammersly Clifford Theorem

If a positive distribution $P(x_1, \ldots, x_n)$ confirms to the pairwise CIs of a UDGM G, then it can be factorized as per the cliques of G. This is the Hammersly Clifford Theorem.

$$P(x_1, \ldots, x_n) \propto \prod_{C \in \mathcal{C}} \psi_C(y_C)$$

The proof of this theorem has been left as an exercise for the reader (Theorem 4.8 of the KF book).

7 Summary

Let P be a distribution and H be an undirected graph of the same set of nodes.

$Factorize(P, H) \implies Global - CI(P, H) \implies Local - CI(P, H) \implies Pairwise - CI(P, H)$

But only for positive distributions, we have the following:

$Pairwise - CI(P, H) \implies Factorize(P, H)$