Class notes
Graphical Models

Amir Navot and Lavi Shpigelman

Winter 2003

1 Introduction

In the HMM discussion we made use of a graphical representation in order to derive independences in the probability. The model that was represented in graphical form enabled us to efficiently learn and make predictions.

In these notes we generalize the use of graphical representation for more complicated models. The graph will be used as easily accessible information regarding the distribution which allows speedup of computations.

We present two different graphical models, the directed and undirected. In these notes we present them and discuss their properties. Later on, the 3 problems that were presented for HMMs (likelihood, inference and learning) will be generalized to these models.

2 Undirected Graphical Models

Definition 1 (conditional independence) *A is independent of B given C (denoted by $(A \perp B)$) if*

$$\begin{aligned} & p(A|B,C) &=& \; p(A|C) \\ \iff & p(B|A,C) &=& \; p(B|C) \\ \iff & p(A,B|C) &=& \; p(A|C)p(B|C) \end{aligned}$$

Proof:[of equivalency]

$$\begin{aligned} p(A|B,C) &= P(A|C) \\ \iff p(A|B,C)p(B|C) &= p(A|C)p(B|C) \end{aligned}$$

$$\begin{aligned}
\iff \quad p(A,B|C) &= p(A|C)p(B|C) \\
\iff \quad \frac{p(A,B|C)}{p(A|C)} &= p(B|C) \\
\iff \quad p(B|A,C) &= p(B|C)
\end{aligned}$$

$\square$

Property 1

$$\{A \cup D\} \perp B|C \Rightarrow A \perp B|C, D \perp B|C$$

Definition 2 (Clique) *A set of vertexes is called a (maximal)* clique *if it is fully connected (and it is not a subset of another fully connected set).*

By clique, we will refer to a maximal clique.

Definition 3 (undirected factorization) *A probability function $p(x)$ is* factorized with respect to a graph G *if*

$$p(x) = \prod_{c \in C} \psi_c(x_c)$$

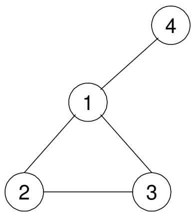
where C is the set of the maximal cliques in G.

The only constraints on $\psi_c(x_c)$ is that

$$\begin{aligned}
\psi_c(x_c) &\geq 0 \\
\sum_x \prod_c \psi_c(x_c) &= 1
\end{aligned}$$

There is no normalization constraint on individual ψs

Example 1 *In the following graph:*



$p(x) = \psi_{123}(x_1, x_2, x_3)\psi_{14}(x_1, x_4)$ is factorized with respect to the graph while $p(x) = \psi_{123}(x_1, x_2, x_3)\psi_{34}(x_3, x_4)$ isn't.

Example 2 *For HMMs, define*

$$\psi_{12} = p(x_1)p(x_2|x_1) \tag{1}$$
$$\psi_{t,t-1} = p(x_t|x_{t-1}) \quad t > 2 \tag{2}$$
$$\psi(x_t, y_t) = p(y_t|x_t) \tag{3}$$

Then

$$p(x, y) = \prod \psi_{t,t-1} \prod \psi(x_t, y_t)$$

Definition 4 (Global Markovity) *A random variable X is* Globally Markov with respect to a graph G *if C separates A and B in the graph $\Longrightarrow A \perp B|C$.*

Definition 5 (Local Markovity) *A random variable X is* Locally Markov with respect to a graph G *if*

$$p(A = a|X \setminus A) = p(A = a|N(A))$$

where $N(A)$ denotes the neighbors of A in G.

Definition 6 (Pairwise Markovity) *A random variable X is* Pairwise Markov with respect to a graph G *if for every non-neighboring pair of vertexes*

$$X_i \perp X_j|\{X \setminus \{X_i, X_j\}\}$$

Statement 1 *def. (4)$\Longrightarrow$ def. (5)$\Longrightarrow$ def. (6) and the opposite does not hold in general but if $\forall x \quad p(x) > 0$ then the 3 definitions are equivalent.*

Statement 2 *If $p(x)$ factorizes with respect to a graph G then X is Globally Markov with respect to G.*

Statement 3 (Hammersley Clifford) *If $\forall x \quad p(x) > 0$ and X is Markov with respect to a graph G then $p(x)$ factorizes with respect to G. (note: $p(x) > 0$ is a sufficient but not necessary condition and can be relaxed).*

Note that if G is a tree there is a locally normalized factorization (i.e. every ψ sums to 1).
The ψs are called *Potentials.*
It is important to realize that conditional independences do not imply unconditional independences.

Statement 4 *If $p(x)$ factorizes with respect to a graph G then X is Locally Markov with respect to G.*

Proof: W.L.G it is enough to prove that

$$p(X_i = s|N(X_i) = 0) = p(X_i = s|X \setminus X_i = 0)$$

We do this by showing that both sides are equal to $\alpha \prod_{c;x_i \in C} \psi_c(s,0)$.

$$\begin{aligned}
p(X_i = s|X \setminus X_i = 0) &= \alpha p(X_i = s, X \setminus X_i = 0) \\
&\overset{\text{factorization}}{=} \underbrace{\alpha \prod_{c:x_i \notin c} \psi_c(0)}_{\alpha_2} \prod_{c:x_i \in c} \psi_c(s,0) \\
&= \alpha_2 \prod_{c:x_i \in c} \psi_c(s,0)
\end{aligned}$$

and

$$\begin{aligned}
p(X_i = s|N(X_i) = 0) &= \alpha p(X_i = s, N(X_i) = 0) \\
&= \alpha \sum_t p\left(()\, X_i = s, N(X-i) = 0, X \setminus \{X_i, N(X_i)\} = t\right. \\
&= \alpha \sum_t \prod_{c:X_i \in C} \psi_c(s,0) \prod_{c:X_i \notin C} \psi_c(t_c,0) \\
&= \alpha_3 \prod_{c:X_i \in C} \psi_c(s,0)
\end{aligned}$$

and $\alpha_3 = \alpha_2$ due to normalization.

□

Proof:[of statement 2] Assume that S separates A and B. then we have to show that

$$p(A|S,B) = p(A|S)$$

Lets define

$$\begin{aligned}
\tilde{A} &= A \cup \{\text{all vertexes that can be reached from A without passing through } S\} \\
\tilde{B} &= X \setminus \{\tilde{A} \cup S\}
\end{aligned}$$

Then $A \subset \tilde{A}$, $B \subset \tilde{B}$ and there are no edges between $\tilde{A}$ and $\tilde{B}$.
Therefore there is no clique that intersects both $\tilde{A}$ and $\tilde{B}$.
Therefore, we can write

$$p(x) \overset{\text{def}}{=} p(\tilde{a}, \tilde{b}, s) = f(\tilde{a}, s) g(\tilde{b}, s)$$

Therefore, $p(x)$ factorizes w.r.t. the graph $\tilde{A} - S - \tilde{B}$.
Therefore, from the previous statement, $p(\tilde{A}|S, \tilde{B}) = p(\tilde{A}|S)$.
Therefore, by property 1 and the fact that $\tilde{A} = A \cup A_2$ and $\tilde{B} = B \cup B_2$, we get the statement.

□

3 Directed Graphical Models

Definition 7 (DAG) *A directed graph is called* acyclic (DAG) *if for all i there are no directed paths from x_i to itself.*

In this context undirected graphs will be DAGs (i.e. acyclic too).

Definition 8 ($\mathbf{p(x)}$ factorizes w.r.t. DAG G) *$p(x)$ factorizes w.r.t. DAG G if*

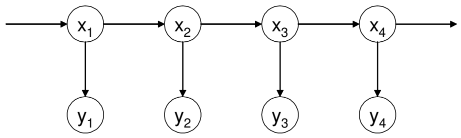
$$p(\mathbf{x}) = \prod_{x_i \in R} p(x_i) \prod_{x_j \notin R} p(x_j | Pa(x_j))$$

where R is the set of G's roots and $Pa(x_j)$ is the set of x_j's parent nodes.

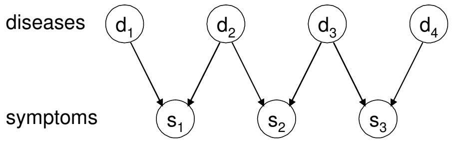
Example 3 (HMM) *For HMM*

$$p(x, y) = p(x_1) \prod_t p(x_t | x_{t-1}) \prod_t p(y_t | x_t)$$

Factorizes w.r.t. the following graph



Example 4 (diseases and symptoms) *in the following graph*



$$p(s, d) = \prod_i p(d_i) \prod_j p(s_j | Pa(s_j))$$

In the directed model, the factors are normalized (i.e. they are probabilities).

Statement 5 *If $p(\mathbf{x})$ factorizes w.r.t. a DAG then all the roots are independent. i.e. $p(x_R) = \prod_{i \in R} p(x_i)$.*

Proof: We prove for the bipartite case (we use the notation from example (4)) and the general case can be shown by induction. (This statement is also proved in corollary 1).

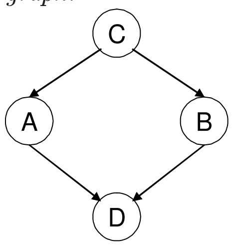
$$
\begin{aligned}
p(\mathbf{d}) &= \sum_{\mathbf{s}} p(\mathbf{d}, \mathbf{s}) \\
&= \sum_{\mathbf{s}} \prod_i p(d_i) \prod_j p(s_j | Pa(s_j)) \\
&= \prod_i p(d_i) \sum_{\mathbf{s}} \prod_j p(s_j | Pa(s_j)) \\
&= \prod_i p(d_i) \sum_{s_1 \ldots s_n} p(s_1 | Pa(s_1)) \ldots p(s_n | Pa(s_n)) \\
&= \prod_i p(d_i) \underbrace{\sum_{s_1} p(s_1 | Pa(s_1))}_{1} \ldots \underbrace{\sum_{s_n} p(s_n | Pa(s_n))}_{1} \\
&= \prod_i p(d_i)
\end{aligned}
$$

$\square$

Definition 9 (directed separation) C directly separates A *and* B *in* G *if every path between* A *and* B *satisfies one of the following:*

1. *The path passes through* $c \in C$ *and the sub-path that includes* c *is not of the form* $\bigcirc \longrightarrow ⓒ \longleftarrow \bigcirc$
2. *The path passes through* $s \notin C$ *(and all the descendents of* s *are also not in* C*) and the sub-path that includes* s *is of the form* $\bigcirc \longrightarrow Ⓢ \longleftarrow \bigcirc$

Example 5 *In the following graph:*



$A \perp B | C$ *because* ACB *satisfies condition 1 and* ABD *satisfies condition 2 but it is not true that* $A \perp B | \{C, D\}$ *since* ADB *does not satisfy condition 1 (or 2).*

Definition 10 (Markovity w.r.t. DAG) $p(x)$ *is* Globally Markov w.r.t. DAG G *if* C *directly separates* A *and* B *in* $G \Longrightarrow A \perp B | C$

An equivalent definition is that for every i, $x_i \perp non\_desc(x_i) | Pa(x_i)$ where non_desc denotes the non-descendant nodes.

Theorem 1 *$p(x)$ is Markov w.r.t. DAG G iff $p(x)$ factorizes w.r.t. G*

Corollary 1 *The roots are independent given the empty set ϕ.*

This follows as ϕ separates the roots.

For a given probability p there are many DAGs that p is factorized w.r.t. them. Some more useful than others. The following statements and definitions elaborate on this subject.

Definition 11 (I-MAP) *Let G be a DAG such that $p(\mathbf{x})$ is Globally Markov w.r.t. G. Then G is an* I-MAP *of $p(\mathbf{x})$.*

I-MAP is short for Independence-MAP.

Statement 6 *Let G be an I-MAP of $p(\mathbf{x})$. Let G' be G with added edges, then G' is also an I-MAP of $p(\mathbf{x})$.*

Proof: Adding an edge to G' can only reduce the set of independences that follow from the Markovity assumption. Therefore G' induces less independences than G and the definition of I-MAP only implies that any dependency that results from the graph is true for the probability (i.e. there can be additional independences in the probability).

□

Note that any full graph over a set of variables implies no dependences and therefore is an I-MAP of any distribution over these variables.

Definition 12 (minimal I-MAP) *Let G be an I-MAP of $p(\mathbf{x})$. If any deletion of an edge from G results in a graph which is not an I-MAP of $p(\mathbf{x})$ then G is a* minimal I-MAP *of $p(\mathbf{x})$.*

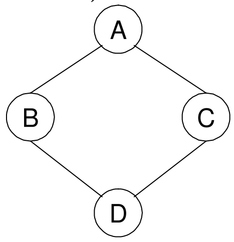
Note that a minimal I-MAP is not unique, nor does it contain a fixed number of edges. In general, a 'good' I-MAP contains a small number of edges since this implies a large number of independences. Unfortunately, the problem of finding an I-MAP with a minimum number of edges is NP-hard

Definition 13 (P-MAP) *Let G be an I-MAP of $p(\mathbf{x})$. If all the independences in $p(\mathbf{x})$ can be derived from G then G is a* P-MAP.

P-MAP is short for Perfect-MAP.

It is clear that we would like to always work with P-MAPS, however, P-MAPs do not exist for some distributions. for example

Example 6 (The rumor problem) *consider the following undirected model*



The independences here are: $A \perp D|B, C$ and $B \perp C|A, D$. It is not possible to construct a P-MAP. This can be seen by explicitly checking all the possible DAGs over this set of vertexes.

Definition 14 (Bayesian Network) *A* Bayesian Network *consists of*

1. *a probability distribution $p(\mathbf{x})$*
2. *an I-MAP of $p(\mathbf{x})$*
3. *additional requirement - the I-MAP is minimal*

Similarly to the case of the undirected graphs, in order to know the probability of a variable we need to know only the values of those variables that separate it from the rest of the graph. For the directed models the definition of this group is a bit more complicated. The following definition and statement characterize it.

Definition 15 (Markov Blanket) *Let $x_i \in G$, then the* Markov Blanket *of x_i is a minimal set of vertexes $MB_i \subset G$ that separates x_i from the rest of G. i.e. $x_i \perp |G \setminus \{MB_i \cup x_i\}$.*

Statement 7 *Let $x_i \in G$, then the Markov Blanket of x_i is*

1. *$Pa(x_i)$*
2. *descendants of x_i*
3. *spouses of x_i (i.e. parents of children of x_i, excluding x_i)*

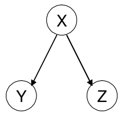
3.1 The D-Separation Algorithm

We would like to find an algorithm that given a DAG G and 3 sets for vertexes A, B, C answers whether $A \perp B|C$ or not. i.e.

$$\text{D-SEP}(A, B|C) = yes \iff A \perp B|C \text{ follows from } G$$

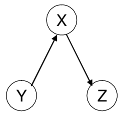
Basically, we will check all paths between A and B to see if they cause dependencies. In order to do so, we define 'active'/'blocked' paths. An *active path* is a path that enables dependency between its end points. A *blocked path* is a path that does not enable dependency. The algorithm will answer 'yes' if all paths from A to B are blocked. To answer whether a path is blocked we look at all threesomes along the path. If a threesome is blocked the whole path is blocked. The possible threesomes are

1. Common cause:

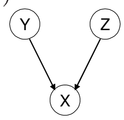
   If $X \in C$ then the threesome is blocked otherwise it is active.

2. Intermediate cause:

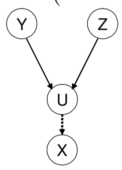
   If $X \in C$ then the threesome is blocked otherwise it is active.

3. Common effect (V structure):

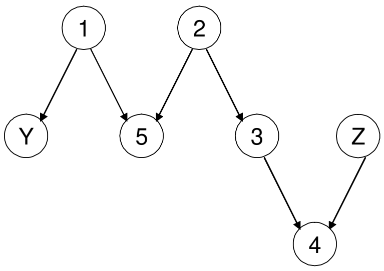
   

   If $X \notin C$ then the threesome is blocked otherwise it is active.

   The same is true if X is a descendant (not a direct child) of Y and Z:

Example 7 *Consider the following graph:*



D-SEP$(Y, Z|5) =$ *yes because the threesome* $\{3, 4, Z\}$ *is blocked.*
D-SEP$(Y, Z|5, 4) =$ *no because all the threesomes are active.*
D-SEP$(Y, Z|5, 4, 2) =$ *yes because the threesome* $\{5, 2, 3\}$ *is blocked.*

In order to check all the paths efficiently we can do a depth-first-search starting from A. Whenever a path is blocked we backtrack. If we reach B then there is an active path. In order to make the algorithm more efficient one can do a pre-step in which all the evidence is floated up to the parents until it hits a V structure.

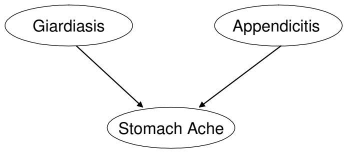
Theorem 2 (Soundness of D-SEP) *If* G *is an I-MAP of* $p(\mathbf{x})$ *then* *D-SEP*$(X, Y|Z) = yes \Longrightarrow X \perp Y|Z$ *in* p.

Theorem 3 (Completeness of D-SEP) *If* G *is an I-MAP of* $p(\mathbf{x})$ *then* *D-SEP*$(X, Y|Z) = no \Longrightarrow$ *there exists a* $p(\mathbf{x})$ *such that* G *is an I-MAP of* $p(\mathbf{x})$ *and* $X \perp Y|Z$ *in* p.

4 Undirected V.S. Directed Graphical Models

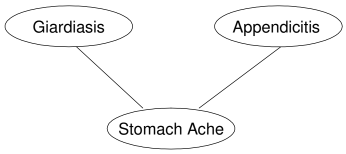
This section deals with the relationship between the directed and undirected models. We will (again) see that they are not equivalent.

Example 8 *The following model*



Says that Giardiasis and Appendicitis are independent but given Stomach Ache they are dependent.

However, the following model



says just the opposite.

Intuitively, if something that happened can be caused by several different causes, then knowing that one of the causes happened reduces the chance of the other causes and vise-versa. This concept is termed *explaining away.* The two most important advantages of the directed model over the undirected is it's capability to model 'explaining away' and to indicate causality.

If a directed graph does not contain V structures then it conveys the exact same information as it's corresponding undirected graph. This is clear from the definitions of directed separation and Markovity.

A consequence of the above observation is that the HMM can be represented as either a directed or undirected graph.

Definition 16 (Directed Tree) *A* directed tree *is a DAG with the additional requirement that any vertex has at most one parent.*

Note that a DAG is a directed tree iff it does not contain a V structure..

Statement 8 *If $p(\mathbf{x})$ is Globally Markov w.r.t. an undirected tree G then it factorizes w.r.t. G (the demand that $p(\mathbf{x}) > 0$ is not required here).*

Proof: Choose a root node at random and construct a directed tree by directing the edges of G according to distance from the root. The resulting graph is a directed tree. Therefore $p(\mathbf{x})$ is Markov w.r.t. the directed tree. Therefore $p(\mathbf{x})$ factorizes as follows

$$p(\mathbf{x}) = p(x_R)\prod_i p(x_i|Pa(x_i))$$

This is also a factorization w.r.t. G because the maximal cliques in G are couples.

□

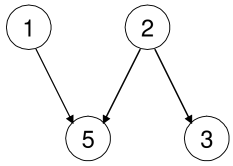
Definition 17 (Moral Graph) *Let G be a DAG. The undirected* Moral Graph *M* of *G is the graph that results in connecting all the nodes that have a common child and deleting the edge directions.*

Statement 9 *If $p(\mathbf{x})$ is Markov w.r.t. a DAG G then it is also Markov w.r.t. the moral graph of G.*

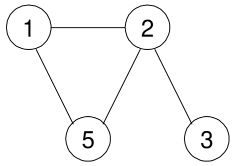
Proof: $p(\mathbf{x}) = \prod_{i\in R} p(x_i)\prod_{i\notin R} p(x_i|Pa(x_i)) = \prod_{c \text{ clique in } M} \psi_c(x_c)$
The last equation is true because when constructing M we connect the parents and children into cliques.

□

Example 9 *Consider G:*



The corresponding moral graph M is



Note that any independence that can be derived from M can be derived from G but there are independences in G that can not be derived from M. e.g. in G $A \perp B$ but not in M. Moralizing a graph reduces independences that can be derived.