Directed and Undirected Graphical Models **2013/2014**

Lecture 4 — October 23rd

Lecturer: Francis Bach *Scribe: Vincent Bodin, Thomas Moreau*

4.1 Notation and probability review

Let us recall a few notations before establishing some properties of directed graphical models. Let $X_1, X_2, \dots, X_n$ be random variables with distribution:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = p_X(x_1, \dots, x_n) = p(x)$$

where x stands for $(x_1, \dots, x_n)$. Given $A \subset \{1, \dots, n\}$, we denote the marginal distribution of x_A by:

$$p(x_A) = \sum_{x \in A^c} p(x_A, x_{A^c}).$$

With this notation we can write the conditional distribution as:

$$p(x_A | x_{A^c}) = \frac{p(x_A, x_{A^c})}{p(x_{A^c})}$$

We also recall the so-called 'chain rule' stating:

$$p(x_1, \dots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1) \dots p(x_n|x_1, \dots, x_{n-1})$$

To end with notations and recalls, we remind the conditional independence characterization:

$$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow p(x, y|z) = p(x|z)\, p(y|z) \Leftrightarrow p(x|y, z) = p(x|z) = \frac{p(x, y|z)}{p(y|z)}$$

4.2 Directed Graphical Model

4.2.1 First definitions and properties

Let $X_1, \dots, X_n$ be n random variables with distribution $p(x) = p_X(x_1, \dots, x_n)$.

Definition 4.1 *Let $G = (V, E)$ be a DAG with $V = \{1, \dots, n\}$. We say that $p(x)$ factorizes in G, denoted $p(x) \in \mathcal{L}(G)$ iff $p(x)$ is of the form:*

$$\forall x, p(x) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}) \text{ such that } f_i \geq 0, \sum_{x_i} f(x_i, x_{\pi_i}) = 1 \tag{4.1}$$

where we recall that π_i stands for the set of parents of the vertex i in G.

We now show that because we assumed that G was a DAG, it implies a particular and convenient form for the f_i above. We have:

Proposition 4.2 *If $p(x) \in \mathcal{L}(G)$ then, for all $i \in \{1, \ldots, n\}$, $f_i(x_i, x_{\pi_i}) = p(x_i|x_{\pi_i})$.*

Proof We prove this by induction on $n = |V|$, the cardinality of the set V. Since G is a DAG, there exists a leaf, *i.e.* a node with no children. Without loss of generality, we can assume that the leaf is labeled by n. We first notice:

$$\begin{aligned}
\forall x, p(x_1, \ldots, x_{n-1}) &= \sum_{x_n} p(x_1, \ldots, x_n) \\
&= \sum_{x_n} \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}) \\
&= \sum_{x_n} f_n(x_n, x_{\pi_n}) \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \\
&= \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \sum_{x_n} f_n(x_n, x_{\pi_n}) \ (*) \\
&= \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \\
&= g(x_1, \ldots, x_{n-1}) \ (**)
\end{aligned} \tag{4.2}$$

The step $(*)$ is justified by the fact that n is a leaf and thus it never appears in any of the π_i for $i \in \{1, \ldots, n-1\}$. Step $(**)$ is also justified by the same kind of reasoning: since n is a leaf it cannot appear in any of the π_i explaining why it is only a function, say g, of $x_1, \ldots, x_{n-1}$. From this result, we can use an induction reasoning noticing that $G - \{n\}$ is still a DAG. To conclude this proof, we simply need to show that, indeed, $f_n(x_n, x_{\pi_n}) = p(x_n|x_{\pi_n})$—this property will automatically propagates by induction. We have:

$$p(x_n, x_{\pi_n}) = \sum_{x_i, i \notin \{n\} \cup \pi_n} p(x) = \left(\sum_{x_i, i \notin \{n\} \cup \pi_n} g(x_1, \ldots, g_{n-1}) \right) f_n(x_n, x_{\pi_n}). \tag{4.3}$$

Noticing that $\sum_{x_i, i \notin \{n\} \cup \pi_n} g(x_1, \ldots, x_{n-1})$ is a function of only x_{π_n}, say $h(x_{\pi_n})$, we can derive:

$$p(x_n|x_{\pi_n}) = \frac{p(x_n, x_{\pi_n})}{\sum_{x'_n} p(x'_n, x_{\pi_n})} = \frac{h(x_{\pi_n}) f_n(x_n, x_{\pi_n})}{h(x_{\pi_n})} = f_n(x_n, x_{\pi_n}). \tag{4.4}$$

■

Hence we can give an equivalent definition for a DAG to the notion of factorization:

Definition 4.3 *(Equivalent definition) $p(x)$ factorizes in G, denoted $p(x) \in \mathcal{L}(G)$ iff:*

$$\forall x, p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}) \tag{4.5}$$

Example 4.2.1

- *(Trivial Graphs) Assume $E = \emptyset$,* i.e. *there is no edges. Then we have $p(x) = \prod_{i=1}^{n} p(x_i)$, implying the random variables $X_1, \ldots, X_n$ are independent. Hence variables are independent if they factorize in the empty graph.*
- *(Complete Graphs) Assume now we have a complete graph (thus with $n(n-1)/2$ edges as we need acyclic for it to be a DAG), we have: $p(x) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1})$, the so-called 'chain rule' which is always true. Every random process factorizes in the complete graph.*

4.2.2 Graphs with three nodes

We give an insight of the different possible behaviors of a graph by thoroughly enumerating the possibilities for a 3-node graph.

- The two first options are the empty graph, leading to independence, and the complete graph that gives no further information than the chain rule.
- (Markov chain) A Markov chain is a certain type of DAG showed in Fig.(4.1). In this configuration we show that we have:

$$p(x, y, z) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Z \mid Y \tag{4.6}$$

Indeed we have:

$$p(z|y, x) = \frac{p(x, y, z)}{p(x, y)} = \frac{p(x, y, z)}{\sum_{z'} p(z', x, y)} = \frac{p(x)p(y|x)p(z|y)}{\sum_{z'} p(x)p(y|x)p(z'|y)} = p(z|y)$$

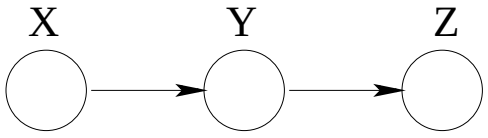


Figure 4.1. Markov Chain

- (Latent cause) It is the type of DAG given in Fig.(4.2). We show that:

$$p(x) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \mid Z \tag{4.7}$$

Indeed:

$$p(x, y|z)\frac{p(x, y, z)}{p(z)} = \frac{p(z)p(y|z)p(x|z)}{p(z)} = p(x|z)p(y|z)$$

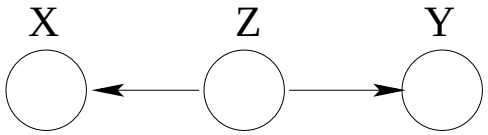


Figure 4.2. Latent cause

- (Explaining away) Represented in Fig.(4.3), we can show for this type of graph:

$$p(x) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \tag{4.8}$$

It basically stems from:

$$p(x,y) = \sum_z p(x,y,z) = p(x)p(y)\sum_z p(z) = p(x)p(y)$$

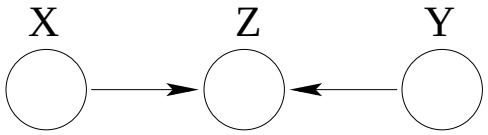


Figure 4.3. Explaining away

Remark 4.2.1 *The use of 'cause' is not advised since observational statistics provide with correlations and no causality notion. Note also that in the 'explaining away' graph, in general* $X \perp\!\!\!\perp Y|Z$ *is not true. Last, it is important to figure that not every relationships can be expressed in terms of graphical modes. As a counter-example take three random variables that are pairwise independent, but not fully independent.*

4.2.3 Inclusion, reversal and marginalization properties

Inclusion property. Here is a quite intuitive proposition about included graphs and their factorization.

Proposition 4.4 *If* $G = (V, E)$ *and* $G' = (V, E')$ *then:*

$$E \subset E' \Leftrightarrow \mathcal{L}(G) \subset \mathcal{L}(G) \tag{4.9}$$

Proof We have $p(x) = \prod_{i=1}^n p(x_i, x_{\pi_i(G)})$. As $E \subset E'$ it is obvious that $\pi_i(G) \subset \pi_i(G')$. Therefore, going back to the definition of graphical models through potential $f_i(x_i, x_{\pi_i})$ we get the result. ■

Reversal property. We also have some reversal properties. Let us first define the notion of V-structure.

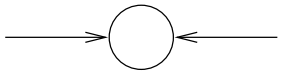


Figure 4.4. V-structure

Definition 4.5 *We say there is a V-structure in $i \in V$ if $|\pi_i| \geq 2$, i.e.i has two or more parents.*

Proposition 4.6 *(Markov equivalence) If $G = (V, E)$ is a DAG and if for $(i, j) \in E, |\pi_i| = 0$ and $|\pi_j| \leq 1$, then (i, j) may be reversed,* i.e. *if $p(x)$ factorizes in G then it factorizes in $G' = (V, E')$ with $E' = (E - \{(i, j)\}) \cup \{(j, i)\}$.*

In terms of 3-nodes graph, this property ensures us that the Markov chain and latent cause are equivalent. On the other hand the V-structure lead to a different class of graph compared to the two others.

Definition 4.7 *An edge (i, j) is said to be covered if $\pi_j = \{i\} \cup \pi_i$.*

By reversing (i, j) we might not get a DAG as it might break the acyclic property. We have the following result:

Proposition 4.8 *Let $G = (V, E)$ be a graph and $(i, j) \in E$ a covered edge. Let $G' = (V, E')$ with $E' = (E - \{(i, j)\}) \cup \{(j, i)\}$, then if G' is a DAG, $\mathcal{L}(G) = \mathcal{L}(G')$.*

Marginalization. The underlying question is to know whether the marginalization of a distribution that factorizes in a graphical model also does. This is true for the marginalization with respect to leaf nodes.

Proposition 4.9 *If i is a leaf, then $p(x_j, j \neq i)$ factorizes in the sub-graph obtained by removing the leaf i.*

Proof It is simply derived by the following computation:

$$\begin{aligned} p(x_1, ...x_{n-1}) &= \sum_{x_n} p(x_1, \ldots, x_n) \\ &= \sum_{x_n} \left(\prod_{i=1}^{n-1} p(x_i|x_{\pi_i}) p(x_n|x_{\pi_n}) \right) \\ &= \prod_{i=1}^{n-1} p(x_i|x_{\pi_i}) \sum_{x_n} p(x_n|x_{\pi_n}) \\ &= \prod_{i=1}^{n-1} p(x_i|x_{\pi_i}) \end{aligned}$$

■

Conditional independence. We finish this section by giving a result that explains that if $p(x)$ factorizes in G then every single random variable is independent from the set of its non-descendants given its parents. From now on, we denote by $\text{nd}(i)$ the set of non-descendants of i.

Proposition 4.10 *If G is a DAG, then:*

$$p(x) \in \mathcal{L}(G) \Leftrightarrow X_i \perp\!\!\!\perp X_{nd(i)} | X_{\pi_i} \tag{4.10}$$

Proof We will only prove the forward implication. Assume $(1, \dots, n)$ is a topological order then:

$$\begin{cases} p(x) &= \prod_{i=1}^{n} p(x_i|x_{\pi_i}) & : \text{because } p(x) \in \mathcal{L}(G) \\ p(x) &= \prod_{i=1}^{n} p(x_i|x_1, \dots, x_{i-1}) & : \text{chain rule, always true} \end{cases}$$

As we chose a topological order, we have $\pi_i \subset \{1, \dots, i-1\}$, and we show by induction that:

$$p(x_i|x_{\pi_i}) = p(x_i|x_1, \dots, x_{i-1}) = p(x_i|x_{\pi_i}, x_{\{1,\dots,i-1\}-\pi_i}).$$

This directly implies that $X_i \perp\!\!\!\perp X_{\pi_i - \{1,\dots,i-1\}} | X_{\pi_i}$. The key idea now is to notice that for all i, there exist a topological order such that $\text{nd}(i) \subset \{1, \dots, i-1\}$. ■

4.2.4 d-separation

We want to answer queries such as, given A, B and C three subsets, is $X_A \perp\!\!\!\perp X_B | X_c$ true? To answer those issues we need the d-separation notion, terminology standing for directed separation. Indeed it is easy to figure out that the notion of separation is not enough in a directed graph and needs to be generalized.

Definition 4.11 *Let $a, b \in V$, a chain from a to b is a sequence of nodes, say $(v_1, \dots, v_n)$ such that $v_1 = a$ and $v_n = b$ and $\forall j, (v_j, v_{j+1}) \in E$ or $(v_j, v_{j+1}) \in E$.*

We can notice that a chain is hence a path in the symmetrized graph, *i.e.* in the graph where if the relation $\rightarrow$ is true then $\leftrightarrow$ is true as well. Assume C is a set that is observed. We want to define a notion of being 'blocked' by this set C in order to answer the underlying question above.

Definition 4.12 *1. A chain from a et b is blocked in d if:*

- *either $d \in C$ and (v_{i-1}, v_i, v_{i+1}) is not a V-structure;*
- *or $d \notin C$ and (v_{i-1}, v_i, v_{i+1}) is a V-structure and no descendants of d is in C.*

2. *A chain from a to b is blocked if and only if it is blocked at any nodes.*

3. *A and B are said to be d-separated by C if and only if all chains that go from $a \in A$ to $b \in B$ are blocked.*

Example 4.2.2

- *(Markov chain) If you try to prove that any set of the future is independent to the past given the present with Markov theory, it might be difficult but the d-separation notion gives the results directly.*

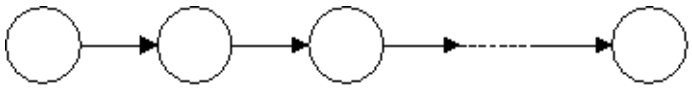


Figure 4.5. Markov chain

- *(Hidden Markov Model) Often used because we only observe a noisy observation of the random process.*

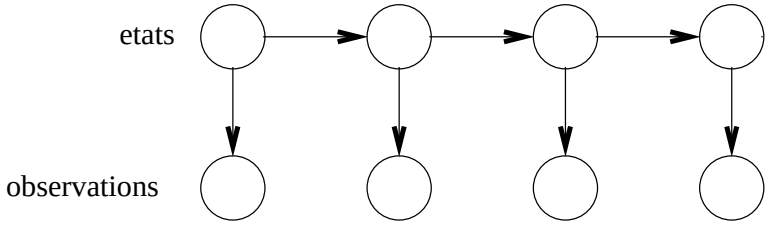


Figure 4.6. Hidden Markov Model

4.3 Undirected graphical models

4.3.1 Definition

Definition 4.13 *Let $G = (V, E)$ be a* **undirected graph**. *We denote by $\mathcal{C}$ a set of cliques of G i.e.a set of sets of fully connected vertices. We say that a probability distribution p factorizes in G and denote $p \in \mathcal{L}(G)$ if $p(x)$ is of the form:*

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \text{ with } \psi_C \geq 0, Z = \sum_x \prod_{C \in \mathcal{C}} \psi_C(x_C).$$

The functions ψ_C are not probability distributions like in the directed graphical models. They are called potentials.

Remark 4.3.1 *With the normalization by Z of this expression, we see that the function ψ_C are defined up to a multiplicative constant.*

Remark 4.3.2 *We may restrict $\mathcal{C}$ to $\mathcal{C}_{max}$, the set of maximal cliques.*

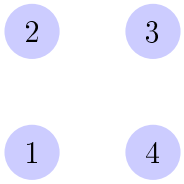
Remark 4.3.3 *This definition can be extended to any function: f is said to factorize in $G \iff f(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C)$.*

4.3.2 Trivial graphs

Empty graphs We consider $G = (V, E)$ with $E = \emptyset$. For $p \in \mathcal{L}(G)$, we get:

$$p(x) = \prod_{i=1}^{n} \psi_i(x_i) \text{ as } \mathcal{C} = \{\{i\} \in V\}$$

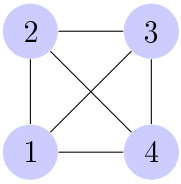
This gives us that $X_1, ..., X_n$ are mutually independent.



Complete graphs We consider $G = (V, E)$ with $\forall i, j \in V, (i, j) \in E$. For $p \in \mathcal{L}(G)$, we get:

$$p(x) = \frac{1}{Z}\psi_V(x_V) \text{ as } \mathcal{C} \text{ is reduced to a single set V}$$

This gives no further information upon the n-sample $X_1, ..., X_n$.



4.3.3 Separation and conditional dependence

Proposition 4.14 *Let $G = (V, E)$ and $G' = (V, E')$ be two undirected graphs.*

$$E \subseteq E' \Rightarrow \mathcal{L}(G) \subseteq \mathcal{L}(G')$$

Definition 4.15 *We say that p satisfies the* **Global Markov property** *w.r.t. G if and only if for all $A, B, S \subset V$ disjoint subsets: A and B are separated by $S \Rightarrow X_A \perp\!\!\!\perp X_B | X_S$.*

Proposition 4.16 *If $p \in \mathcal{L}(G)$ then, p satisfies the Global Markov property w.r.t. G.*

Proof We can consider that $A \cup B \cup S = V$ without loss of generality as we could replace A and B by :

$$\widetilde{A} = A \cup \{a \in V \ / \text{ a and A are not separated by S }\}$$
$$\widetilde{B} = V \setminus \{S \cup \widetilde{A}\}$$

$\widetilde{A}$ and $\widetilde{B}$ are separated by S.

We consider $C \in \mathcal{C}$. It is not possible to have $C \cap A \neq \emptyset$ and $C \cap B \neq \emptyset$ as A and B are separated by S. Then $C \subset A \cup S$ or $C \subset B \cup S$. Let $\mathcal{D}$ be the set of cliques C such that $C \subset A \cup S$ and $\mathcal{D}'$ the set of all other cliques. We have:

$$p(x) = \frac{1}{Z} \prod_{\substack{C \in \mathcal{C} \\ C \subset A \cup S}} \psi_C(x_C) \prod_{C \in \mathcal{D}'} \psi_C(x_C) = f(x_{A \cup S}) g(x_{B \cup S})$$

$$\begin{aligned} p(x_A | x_S, x_B) &= \frac{p(x_A, x_B, x_S)}{p(x_S, x_B)} = \frac{f(x_A, x_S) g(x_S, w_B)}{\sum_{x'_A} f(x'_A, x_S) g(x_B, x_S)} \\ &= \frac{f(x_A, x_S)}{\sum_{x'_A} f(x'_A, x_S)} \\ &= p(x_A | x_S) \end{aligned}$$

Thus, $X_A \perp\!\!\!\perp X_B | X_S$. ■

Théorème 4.17 *(Hammersley - Clifford) If* $\forall x,\ p(x) > 0$ *then* $p \in \mathcal{L}(G) \iff$ *p satisfies the global Markov property.*

4.3.4 Marginalization

As for directed graphical models, we also have a marginalization notion in undirected graphs. It is slightly different. If $p(x)$ factorizes in G, then $p(x_1, \ldots, x_n)$ factorizes in the graph where the node n is removed and all neighbors are connected.

Proposition 4.18 *Let* $G = (V, E)$ *be an undirected graph and* $G' = (V', E')$ *the graph obtain for* $V' = V \setminus \{n\}$ *and* E' *the set of edge connecting all the neighbors of n. If* $p \in \mathcal{L}(G)$ *then* $p(x_1, ..., x_{n-1}) \in \mathcal{L}(G')$.

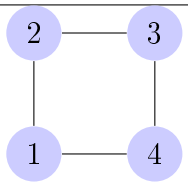
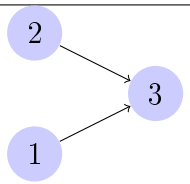
We now introduce the notion of Markov blanket

Definition 4.19 *For* $i \in V$, *the* **Markov blanket** *is the smallest set of nodes that makes* X_i *independent to the rest of the graph.*

Remark 4.3.4 *The Markov blanket in an undirected graph for* $i \in V$ *is the set of its neighbors. For a directed graph, it is the union of all parents, all children and parents of children.*

4.3.5 Relation between directed and undirected graphical models

Since now we have seen that many notions developed for directed graph naturally extended to undirected graphs. The raising question is thus to know whether we can find a theory including both directed and undirected graphs, in particular, is there a way—for instance by symmetrizing the directed graph as we have done repeatedly—to find a general equivalence between those two notions. The answer is no, as we will discuss—though it might work in some special cases described above.

	Directed graphical model	Undirected graphical model
Factorization	$p(x) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})$	$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$
Set independence	d-separation	separation
Marginalization	non close	close
Not equivalent		

Let G be DAG. Can we find G' undirected such that $\mathcal{L}(G) = \mathcal{L}(G')$? $\mathcal{L}(G) \subset \mathcal{L}(G')$?

Definition 4.20 *Let $G = (V, E)$ be a* DAG. *The* **symmetrized graph** *of G is $\tilde{G} = (V, \tilde{E})$, with $\tilde{E} = \{(u, v), (v, u) / (u, v) \in E\}$.*

Definition 4.21 *Let $G = (V, E)$ be a* DAG. *The* **moralized graph** *$\bar{G}$ of G is the symmetrized graph $\tilde{G}$, where we add edge such that for all $v \in V$, π_v is a clique.*

We admit the following proposition:

Proposition 4.22 *Let G be a* DAG *without any V-structure, then $\bar{G} = \tilde{G}$ and $\mathcal{L}(G) = \mathcal{L}(\tilde{G}) = \mathcal{L}(\bar{G})$.*

In case there is a V-structure in the graph, we can only conclude:

Proposition 4.23 *Let G be a* DAG, *then $\mathcal{L}(G) \subset \mathcal{L}(\bar{G})$.*

$\bar{G}$ is minimal for the number of edges in the set H of undirected graphs such that $\mathcal{L}(G) \subset \mathcal{L}(H)$.

Not all conditional independence structure for random variables can be factorized in a graphical model (directed or undirected).