Directed and undirected graphical models **Fall 2017**

Lecture 4 — October 18th

Lecturer: Guillaume Obozinski *Scribe:*

In this lecture, we will assume that all random variables are discrete, to keep notations as simple as possible. All the theory presented generalizes immediately to continuous random variables that have a density by replacing

- the discrete probability distributions considered in this lecture by densities
- summations by integration with respect to a measure of reference (most of the time the Lebesgue measure).

4.1 Notation and probability review

4.1.1 Notations

We review some 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:

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

4.1.2 Independence and conditional independence

Let A, B, and C be disjoint.
We will say that X_A is (marginally) independent of X_B and write $X_A \perp\!\!\!\perp X_B$ if and only if

$$p(x_A, x_B) = p(x_A)\, p(x_B) \qquad \forall (x_A, x_B), \tag{4.1}$$

or equivalently if and only if

$$p(x_A|x_B) = p(x_A) \qquad \forall x_A, x_B \quad \text{s.t.} \quad p(x_B) > 0. \tag{4.2}$$

Similarly we will say that X_A is independent from X_B conditionally on X_C (or given X_C) and we will write $X_A \perp\!\!\!\perp X_B \mid X_C$ if and only if

$$p(x_A, x_B|x_C) = p(x_A|x_C)\, p(x_B|x_C) \qquad \forall x_A, x_B, x_C \quad \text{s.t.} \quad p(x_C) > 0, \tag{4.3}$$

or equivalently if and only if

$$p(x_A|x_B, x_C) = p(x_A|x_C) \qquad \forall x_A, x_B, x_C \quad \text{s.t.} \quad p(x_B, x_C) > 0. \tag{4.4}$$

More generally we will say that the $(X_{A_i})_{1\leq i\leq k}$ are *mutually independent* if and only if

$$p(x_{A_1}, \dots, x_{A_k}) = \prod_{i=1}^{k} p(x_{A_i}) \quad \forall x_{A_1}, \dots, x_{A_k},$$

and that they are *mutually independent conditionally on* X_C (or given X_C) if and only if

$$p(x_{A_1}, \dots, x_{A_k}|x_C) = \prod_{i=1}^{k} p(x_{A_i}|x_C) \qquad \forall x_{A_1}, \dots, x_{A_k}, x_C \quad \text{s.t.} \quad p(x_C) > 0.$$

Remark 4.1.1 *Note that the conditional probability $p(x_A, x_B|x_C)$ is the probability distribution over (X_A, X_B) if X_C is known to be equal to x_C. In practice, it means that if the value of X_C is* observed *(e.g. via a measurement) then the distribution over (X_A, X_B) is $p(x_A, x_B|x_C)$. The conditional independence statement $X_A \perp\!\!\!\perp X_B \mid X_C$ should therefore be interpreted as "when the value of X_C is observed (or given) X_A and X_B are independent".*

Remark 4.1.2 *(Pairwise independence vs mutual independence) Consider a collection of r.v. $(X_1, \dots, X_n)$. We say that these variables are pairwise independent if for all $1 \leq i < j \leq n, \quad X_i \perp\!\!\!\perp X_j$. Note that this is different than assuming that $X_1, \dots, X_n$ are mutually (or jointly or globally) independent. A standard counter-example is as follows: given two variables X, Y that are independent coin flips define Z via the XOR function $\oplus$ with $Z = X \oplus Y$. Then, the three random variables X, Y, Z are* pairwise independent*, but not* mutually independent*. (Prove this as an exercise.) The notations presented for* pairwise independence *could be generalized to collections of variables that are* mutually independent*.*

Three Facts About Conditional Independence

1. **Can repeat variables:** $X \perp\!\!\!\perp (Y, Z) \mid Z, W$ is the same as $(X, Z) \perp\!\!\!\perp Y \mid Z, W$. The repetition is redundant but may be convenient notation.
2. **Decomposition:** $X \perp\!\!\!\perp (Y, Z) \mid W$ implies that $X \perp\!\!\!\perp Y \mid W$ and $X \perp\!\!\!\perp Z \mid W$.
3. The chain rule applies to conditional distributions:

$$p(x, y|z) = p(x|y, z)\, p(y|z). \tag{4.5}$$

Proving each of these three facts are good simple exercises.

4.2 Directed Graphical Model

Graphical models combine probability and graph theory into an efficient data structure. We want to be able to handle probabilistic models of hundreds of variables. For example, assume we are trying to model the probability of diseases given the symptoms, as shown below.

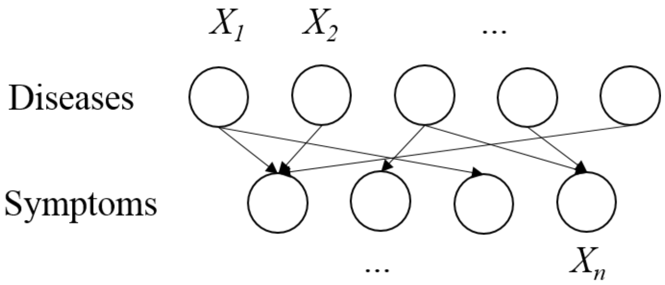


Figure 4.1. Nodes representing binary variables indicating the presence or not of a disease or a symptom.

We have n nodes, each a binary variable ($X_i \in \{0, 1\}$), indicating the presence or absence of a disease or a symptom. The number of joint probability terms would grow exponentially. For 100 diseases and symptoms, we would need a table of size 2^{100} to store all the possible states. This is clearly intractable. Instead, we will use graphical models to represent the relationships between nodes.

General issues in this class

1. Representation → DGM, UGM / parameterization → exponential family
2. Inference (computing $p(x_A|x_B)$) → sum-product algorithm
3. Statistical estimation → maximum likelihood, maximum entropy

A directed graphical model, also historically called "Bayesian network" when the variables are discrete, represents a *family of distributions*, denoted $\mathcal{L}(G)$, where $\mathcal{L}(G) \triangleq \{p : \exists$ legal factors, f_i, s.t. $p(x_V) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i})\}$, where the legal factors satisfy $f_i \geq 0$ and $\sum_{x_i} f_i(x_i, x_{\pi_i}) = 1 \ \forall i, x_{\pi_i}$.

4.2.1 First definitions and properties

Let $X_1, \dots, X_n$ be n random variables with joint 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 $p(x) \in \mathcal{L}(G)$, if there exists some functions f_i, called factors, such that:*

$$\begin{aligned} \forall x,\ p(x) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}) \\ f_i \geq 0, \quad \forall i, \forall x_{\pi_i}, \sum_{x_i} f_i(x_i, x_{\pi_i}) = 1 \end{aligned} \tag{4.6}$$

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

We prove the following useful and fundamental property of directed graphical models: if a probability distribution factorizes according to a directed graph $G = (V, E)$, the distribution obtained by marginalizing a leaf[1] i factorizes according to the graph induced on $V \backslash \{i\}$.

Proposition 4.2 (Leaf marginalization) *Suppose that p factorizes in G, i.e. $p(x_V) = \prod_{j=1}^{n} f_j(x_j, x_{\pi_j})$. Then for any leaf i, we have that $p(x_{V\backslash\{i\}}) = \prod_{j \neq i} f_j(x_j, x_{\pi_j})$, hence $p(x_{V\backslash\{i\}})$ factorizes in $G' = (V\backslash\{i\}, E')$, the induced graph on $V\backslash\{i\}$.*

Proof Without loss of generality, we can assume that the leaf is indexed by n. Since it is a leaf, we clearly have that $n \notin \pi_i, \forall i \leq n-1$. We have the following computation:

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

■

Remark 4.2.1 *Note that the new graph obtained by removing a leaf is still a DAG. Indeed, since we only removed edges and nodes, if there was a cycle in the induced graph, the same cycle would be present in the original graph, which is not possible since it is DAG.*

Remark 4.2.2 *Also, by induction, this result shows that in the definition of factorization we do not need to assume that p is a probability distribution. Indeed, if any function p satisfies (4.6) then it is a probability distribution, because its non-negative as a product of non-negative factors and it sums to 1 by using formula proved by induction.*

[1] We call here a *leaf* or *terminal node* of a DAG a node that has no descendant.

Lemme 4.3 *Let A, B, C be three sets of nodes such that $C \subset B$ and $A \cap B = \varnothing$. If $p(x_A \mid x_B)$ is a function of only (x_A, x_C) then $p(x_A|x_B) = p(x_A|x_C)$.*

Proof We denote by $f(x_A, x_C) := p(x_A|x_B)$ the corresponding function. Then $p(x_A, x_B) = p(x_A \mid x_B)\, p(x_B) = f(x_A, x_C)\, p(x_B)$. By summing over $x_{B\setminus C}$, we have

$$p(x_A, x_C) = \sum_{x_{B\setminus C}} p(x_A, x_B) = f(x_A, x_C) \sum_{x_{B\setminus C}} p(x_B) = f(x_A, x_C)\, p(x_C),$$

which proves that $p(x_A|x_C) = f(x_A, x_C) = p(x_A|x_B)$. ■

Now we try to characterize the factor functions. The following result will imply that if p factorizes in G, then we have a uniqueness of the factors.

Proposition 4.4 *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 Assume, without loss of generality, that the nodes are sorted in a topological order. Consider a node i. Since the node are in topological order, for any $1 \leq j \leq n$, we have $\pi_j \subset \{1, \ldots, j-1\}$; as a consequence we can apply Proposition 4.2 $n-i$ times to obtain that $p(x_1, \ldots, x_i) = \prod_{j \leq i} f(x_j, x_{\pi_j})$. Since we also have $p(x_1, \ldots, x_{i-1}) = \prod_{j<i} f(x_j, x_{\pi_j})$, taking the ratio, we have

$$p(x_i \mid x_1, \ldots, x_{i-1}) = f(x_i, x_{\pi_i}).$$

Since $\pi_i \subset \{1, \ldots, i-1\}$, this entails by the previous lemma that $p(x_i \mid x_1, \ldots, x_{i-1}) = p(x_i \mid x_{\pi_i}) = f(x_i, x_{\pi_i})$. ■

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

Definition 4.5 *(Equivalent definition) The probability distribution $p(x)$ factorizes in G, denoted $p(x) \in \mathcal{L}(G)$, iff*

$$\forall x, \quad p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}) \tag{4.7}$$

Example 4.2.1

- *(Trivial Graphs) Assume $E = \varnothing$,* i.e. *there are no edges. We then have $p(x) = \prod_{i=1}^{n} p(x_i)$, implying the random variables $X_1, \ldots, X_n$ are independent, that is variables are mutually 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 acyclicity 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 probability distribution factorizes in complete graphs. Note that there are $n!$ complete graph possible that are all equivalent...*
- *(Graphs with several connected components) If G has several connected components $C_1, \ldots, C_k$, then $p \in \mathcal{L}(G) \;\Rightarrow\; p(x) = \prod_{j=1}^{k} p(x_{C_j})$* (Exercise). *As a consequence, each connected component can be treated separately. In the rest of the lecture, we will therefore focus on connected graphs.*

4.2.2 Graphs with three nodes

We consider all connected graphs with three nodes, except for the complete graph, which we have already discussed.

- (Markov chain) The Markov chain on three nodes is illustrated on Fig.(4.2). For this graph we have

$$p(x, y, z) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \mid Z \tag{4.8}$$

  Indeed we have:

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

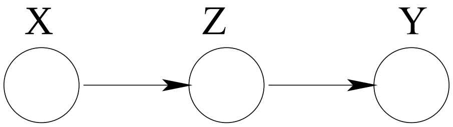


Figure 4.2. Markov Chain

- (Latent cause) It is the type of DAG given in Fig.(4.3). We show that:

$$p(x, y, z) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \mid Z \tag{4.9}$$

  Indeed:

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

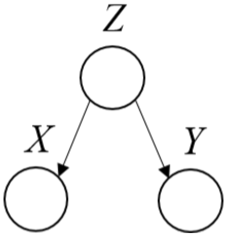


Figure 4.3. Common latent "cause"

- (Explaining away) Represented in Fig.(4.4), we can show for this type of graph:

$$p(x) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \tag{4.10}$$

  It basically stems from:

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

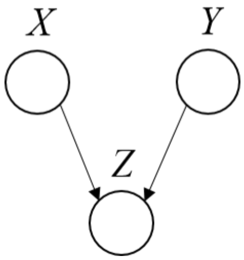


Figure 4.4. Explaining away, or v-structure

Remark 4.2.3 *The word "cause" should here be between quotes and used very carefully, because the same way that* Correlation is not causation*, conditional dependance is not causation either. This is however the historical name for this model. The reason why cause is a bad name, and that* latent factor *might be better, is that the factorisation properties that are encoded by graphical models do not in general correspond to the existence of a causal mechanisms, but only to conditional independence relations.*

Remark 4.2.4 *If p factorizes in the "latent cause" graph, then $p(x, y, z) = p(z)p(x|z)p(y|z)$. But using Bayes' rule $p(z)\,p(x|z) = p(x)\,p(z|x)$, adn so we also have that $p(x, y, z) = p(x)p(z|x)p(y|z)$ which shows that p is a Markov chain (i.e. factorizes in the Markov chain graph). This is an example of* basic edge reversal *that we will discuss in the next section. Note that we proceeded by equivalence, which shows that the Markov chain graph and the "latent cause" graph and the reversed Markov chain graph are in fact equivalent in the sense that distribution that factorize according to one factorize according to the others. This is what we will call Markov equivalence.*

Remark 4.2.5 *In the "explaining away" graph, in general $X \perp\!\!\!\perp Y|Z$ is not true in the sense that there exist elements in $\mathcal{L}(G)$ such that this statement is violated.*

Remark 4.2.6 *For a graph, ($p \in \mathcal{L}(G)$) implies that p satisfies some list of (positive) conditional independence statements (CIS). The fact that p is in $\mathcal{L}(G)$ cannot guarantee that a given CIS does not hold. This should be obvious because the independent distribution belongs to all graphical models and satisfies all CIS...*

Remark 4.2.7 *It is important to note that not all lists of CIS correspond to a graph, in the sense that there are lists of CIS for which there exists is no graph such that $\mathcal{L}(G)$ is formed exactly of the distributions which satisfy only the conditional independences that are listed or that are consequences of the ones listed. In particular there is no graph G on three variables such that $\mathcal{L}(G)$ contains all distributions on (X,Y,Z) that satisfy $X \perp\!\!\!\perp Y$, $Y \perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Z$ and does not contain distributions for which any of these statements is violated. (Remember that pairwise independence does not imply mutual independence: see Remark 4.1.2).*

4.2.3 Inclusion, reversal and marginalization properties

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

Proposition 4.6 *If $G = (V, E)$ and $G' = (V, E')$ then:*

$$E \subset E' \quad \Rightarrow \quad \mathcal{L}(G) \subset \mathcal{L}(G'). \tag{4.11}$$

Proof If $p \in \mathcal{L}(G)$, then $p(x) = \prod_{i=1}^{n} p(x_i, x_{\pi_i(G)})$. Since $E \subset E'$, it is obvious that $\pi_i(G) \subset \pi_i(G')$, and we can define $f_i(x_i, x_{\pi_i(G')}) := p(x_i | x_{\pi_i(G)})$. Since $p(x) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i(G')})$ and f_i meets the requirements of Definition 4.1, this proves that $p \in \mathcal{L}(G')$. ■

The converse of the previous proposition is not true. In particular, different graphs can define the same set of distribution. We introduce first some new definitions:

Definition 4.7 *(Markov equivalence) We say that two graphs G and G' are Markov equivalent if $\mathcal{L}(G) = \mathcal{L}(G')$.*

Proposition 4.8 *(Basic edge reversal) If $G = (V, E)$ is a DAG and if for $(i, j) \in E$, i has no parents and the only parent of j is i, then the graph obtained by reversing the edge (i, j) is Markov equivalent to G.*

Proof First, note that by reversing such an edge no cycle can be created because the cycle would necessarily contain (j, i) and j has no parent other than i. Using Bayes' rule: $p(x_i)\, p(x_j | x_i) = p(x_j)\, p(x_i | x_j)$ we convert the factorization w.r.t. to G to factorization w.r.t. to the graph obtained by edge reversal. ■

Informally, the previous result can be reformulated as: an edge reversal that does not remove or creates any v-structure leads to a graph which is Markov equivalent.

When applied to the 3-nodes graphs considered earlier, this property proves that the Markov chain and the "latent cause" graph are equivalent. On the other hand, the fact that the "explain away" graph has a v-structure is the reason why it is not equivalent to the others.

Definition 4.9 *(Covered edge) An edge (i, j) is said to be covered if $\pi_j = \{i\} \cup \pi_i$.*

Proposition 4.10 *(Covered edge reversal) Let $G = (V, E)$ be a DAG and $(i, j) \in E$ a covered edge. Let $G' = (V, E')$ with $E' = (E \backslash \{(i, j)\}) \cup \{(j, i)\}$, then G' is necessarily also a DAG and $\mathcal{L}(G) = \mathcal{L}(G')$.*

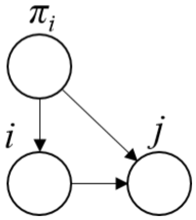


Figure 4.5. Edge (i, j) is covered

Proof Exercise in Homework 2. ■

Marginalization. We have proved in Proposition 4.2 that if $p(x_1, \dots, x_n)$ factorizes in G, the distribution obtained by marginalizing a leaf i factorizes in the graph G' induced on $V \backslash \{i\}$ by G. A nice property of the obtained graph is that all the conditional independences between variables $X_1, \dots, X_{n-1}$ that were implied by G are still implied by G': marginalization has lost CI information about X_n but not about the rest of the distribution. It would be natural to try to generalize this and a legitimate question is: if we marginalise a node i in a distribution of $\mathcal{L}(G)$ is there a simple construction of a graph G' such that the marginalized distribution factorizes in G' and such that all the CIS that hold in G and do not involve X_i are still implied by G'. Unfortunately this is not true. Another less ambitious natural question is then: is there a unique smallest graph G' such that if $p \in \mathcal{L}(G)$ then the distribution obtained by marginalizing i is in $\mathcal{L}(G')$. Unfortunately this is not the case either, as illustrated by the following exemple.

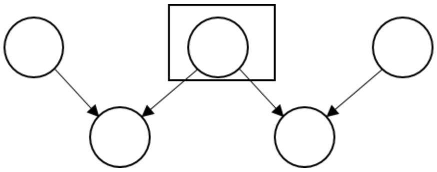


Figure 4.6. Marginalizing the boxed node would not result in family of distributions that cannot be exactly represented by a directed graphical model and one can check that there is no unique smallest graph in which the obtained distribution factorize.

Conditional independence with the non-descendents. In a Markov chain, a well known property is that the X_t is independent of the past given X_{t-1}. This result generalizes as follows in a directed graphical model: if $p(x)$ factorizes in G then every single random variable is independent from the set of its non-descendants given its parents.

Definition 4.11 *The set of non-descendants of i denoted $nd(i)$ is the set of nodes that are not descendants of i.*

Lemme 4.12 *For a graph $G = (V, E)$ and a node i, there exists a topological order such that all elements of $nd(i)$ appear before i.*

Proof This is easily proved constructively: we construct the topological order in reverse order. At each iteration we remove a node among leaves (of the remaining graph) which we add in the reverse order, and specifically, if some leaves are descendants of i then we remove one of those. If at any iteration there is no leaf that is a descendant of i, it means that all descendants of i have been removed from the graph. Indeed, if there were some descendants of i left in the graph, since all their descendants are descendants of i as well there would exist a leaf node which is a descendant of i. This procedure thus removes all strict descendants of i first, then i and then only all elements of $nd(i)$. ■

With this lemma, we can show our main result.

Proposition 4.13 *If G is a DAG, then:*

$$p(x) \in \mathcal{L}(G) \Leftrightarrow X_i \perp\!\!\!\perp X_{nd(i)} | X_{\pi_i} \tag{4.12}$$

Proof First, we consider the $\Rightarrow$ direction. Based on the previous lemma we can find an order such that $nd(i) = \{1, \dots, i-1\}$. But we have proven in Proposition 4.4 that $p(x_i|x_{\pi_i}) = p(x_i|x_{1:(i-1)})$, which given the order chosen is also $p(x_i|x_{1:(i-1)}) = p(x_i|x_{\pi_i}, x_{nd(i)\backslash\pi_i})$; this proves what we wanted to show: $X_i \perp\!\!\!\perp X_{nd(i)\backslash\pi_i} | X_{\pi_i}$.

We now prove the $\Leftarrow$ direction. Let $1:n$ be a topological order, Then $\{1, \cdots i{-}1\} \subseteq nd(i)$. (By contradiction, suppose $j \in \{1 \cdots i-1\}$ and $j \notin nd(i)$, then $\exists$ path from i to j, which contradicts the topological order property as there would be an edge from i to an element of $\{1, \dots, i-1\}$.)

By the chain rule, we always have $p(x_V) = \prod_{i=1}^{n} p(x_i|x_{1:i-1})$ but by the conditional independence assumptions $p(x_i|x_{1:i-1}) = p(x_i|x_{\pi_i})$, hence the result by substitution. ■

4.2.4 d-separation

Given a graph G and A, B and C, three subsets it would be useful to be able to answer the question: is $X_A \perp\!\!\!\perp X_B \mid X_C$ true for all $p \in \mathcal{L}(G)$? An answer is provided by the concept of d-separation, or directed separation.

We call a chain a path in the symmetrized graph, *i.e.* in the graph the undirected graph obtained by ignoring the directionality of the edges.

Definition 4.14 *(Chain) 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 $(v_{j+1}, v_j) \in E$.*

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 4.15 *(Blocking node in a chain, blocked chain and d-separation)*

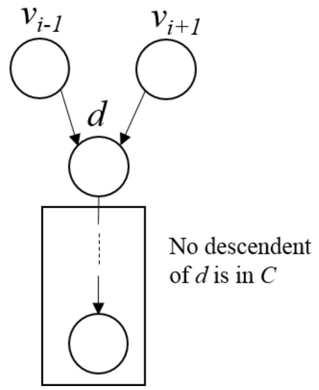


Figure 4.7. D-separation

1. *A chain from a et b is blocked at d if:*
   - *either $d \in C$ and (v_{i-1}, d, v_{i+1}) is not a v-structure;*
   - *or $d \notin C$ and (v_{i-1}, d, 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 node.*
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:*** *Applying d-separation to the Markov chain retrieves the well know results that the future is independent to the past given the present.*

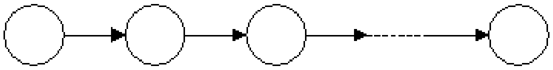


Figure 4.8. Markov chain

- ***Hidden Markov Model:*** *We can apply it as well to the hidden Markov chain graph of Figure 4.9.*

4.2.5 Bayes ball algorithm

Checking whether two nodes are d-separated is not always easy. The Bayes ball algorithm is an intuitive "reachability" algorithm to answer this question. Suppose we want to determine if X is conditionally independent from Z given Y. The principle of the algorithm is to place initially a ball on each of the nodes in X, to then let them bounce around according to some

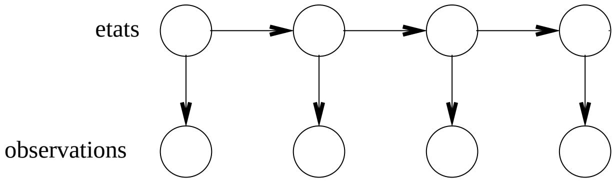


Figure 4.9. Hidden Markov Model

rules described below and to see if any reaches Z. $X \perp\!\!\!\perp Z \mid Y$ is true if none reached Z, but not otherwise.

The rules are as follows for the three canonical graph structures. Note that the balls are allowed to travel in either direction along the edges of the graph.

1. **Markov chain:** Balls pass through when we do not observe Y, but are blocked otherwise.

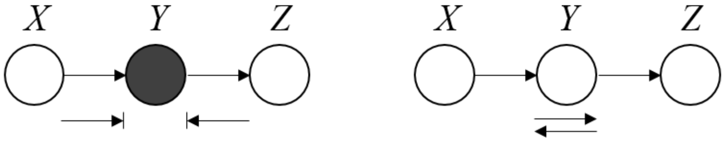


Figure 4.10. Markov chain rule: When Y is observed, balls are blocked (left). When Y is not observed, balls pass through (right)

2. **Two children:** Balls pass through when we do not observe Y, but are blocked otherwise.

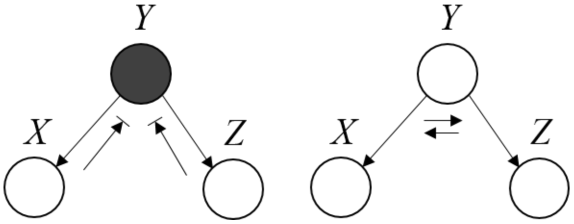


Figure 4.11. Rule when X and Z are Y's children: When Y is observed, balls are blocked (left). When Y is not observed, balls pass through (right)

3. **v-structure:** Balls pass through when we observe Y, but are blocked otherwise.

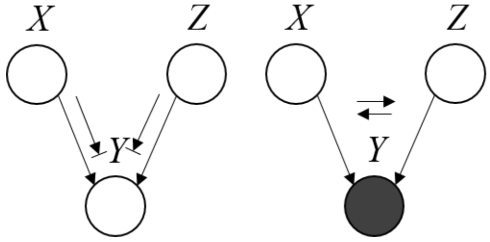


Figure 4.12. v-structure rule: When Y is not observed, balls are blocked (left). When Y is observed, balls pass through (right)

4.3 Undirected graphical models

4.3.1 Definition

Definition 4.16 *Let $G = (V, E)$ be a* **undirected graph***. We denote by $\mathcal{C}$ the set of all cliques of G . We say that a probability distribution p factorizes in G and write $p \in \mathcal{L}(G)$ if p(x) is of the form:*

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

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.*

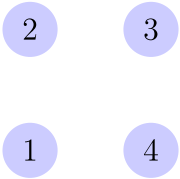
Remark 4.3.3 *This definition can be extended to any function: f is said to factorize in $G \iff f(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C)$.*

4.3.2 Trivial graphs

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

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

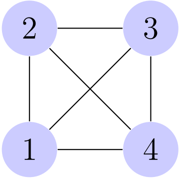
So $X_1, ..., X_n$ must be mutually independent.



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:

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

This places no constraints on the distribution of $(X_1, ..., X_n)$.



4.3.3 Separation and conditional dependence

Proposition 4.17 *Let $G = (V, E)$ and $G' = (V, E')$ be two undirected graphs.*

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

Proof The cliques of G are included in cliques of G'. ■

Definition 4.18 *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) $\Rightarrow$ ($X_A \perp\!\!\!\perp X_B \mid X_S$).*

Proposition 4.19 *If $p \in \mathcal{L}(G)$ then, p satisfies the Global Markov property w.r.t. G.*

Proof We suppose without loss of generality that A, B, and S are disjoint sets such that $A \cup B \cup S = V$, as we could otherwise replace A and B by :

$$A' = A \cup \{a \in V/a \ \mathit{and}\ A\ \mathit{are\ not\ separated\ by}\ S\}$$

$$B' = V \setminus \{S \cup A'\}$$

A' and B' are separated by S and we have the disjoint union $A' \cup B' \cup S = V$. If we can show that $X_{A'} \perp\!\!\!\perp X_{B'} | X_S$, then by the decomposition property, we also have that $X_A \perp\!\!\!\perp X_B | X_S$ for any subset A of A' and B of B', giving the required general case.

We consider $C \in \mathcal{C}$. It is not possible to have both $C \cap A \neq \emptyset$ and $C \cap B \neq \emptyset$ as A and B are separated by S and C is a clique. Thus $C \subset A \cup S$ or $C \subset B \cup S$ (or both if $C \subset S$). 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:

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

Thus:

$$p(x_A, x_S) = \frac{1}{Z} f(x_A, x_S) \sum_{x_B} g(x_B, x_S) \implies p(x_A | x_S) = \frac{f(x_A, x_S)}{\sum_{x'_A} f(x'_A, x_S)}.$$

Similarly: $p(x_B|x_S) = \frac{g(x_B, x_S)}{\sum_{x'_B} g(x'_A, x_S)}$. Hence:

$$p(x_A, x_S) p(x_B | x_S) = \frac{\frac{1}{Z} f(x_A, x_S) g(x_B, x_S)}{\frac{1}{Z} \sum_{x'_A} f(x'_A, x_S) \sum_{x'_B} g(x'_A, x_S)} = \frac{p(x_A, x_B, x_S)}{p(x_S)} = p(x_A, x_B | x_S).$$

i.e. $X_A \perp\!\!\!\perp X_B | X_S$. ■

Theorem 4.20 *(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 $p(x_1, \ldots, x_{n-1})$ factorizes in the graph where the node n is removed and all neighbors are connected.

Proposition 4.21 *Let* $G = (V, E)$ *be an undirected graph. Let* $G' = (V', E')$ *be the graph where* n *is removed and its neighbors are connected, i.e.* $V' = V \setminus \{n\}$*, and* E' *is obtained from the set* E *by first connecting together all the neighbours of* n *and then removing* n*. If* $p \in \mathcal{L}(G)$ *then* $p(x_1, ..., x_{n-1}) \in \mathcal{L}(G')$*. Hence undirected graphical models are closed under marginalization as the construction above is true for any vertex.*

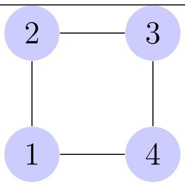
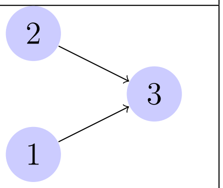
We now introduce the notion of Markov blanket

Definition 4.22 *For* $i \in V$*, the* **Markov blanket** *of a graph* G *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	$p(x) = \prod_{i=1}^{n} p(x_i \vert x_{\pi_i})$	$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$
Set independence	d-separation $[x_i \perp\!\!\!\perp x_{nd(i)} \vert x_{\pi_i}]$ (and many more)	separation $[X_A \perp\!\!\!\perp X_B \vert X_S]$
Marginalization	not closed in general, only when marginalizing leaf nodes	closed
Difference	 grid	 v-structure

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 4.23 *Let $G = (V, E)$ be a* DAG*. The* **symmetrized graph** *of G is $\tilde{G} = (V, \tilde{E})$, with $\tilde{E} = \{(u, v), (v, u) / (u, v) \in E\}$, ie. an edge going the opposite direction is added for every edge in E.*

Definition 4.24 *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 4.25 *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 4.26 *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).