Bayesian Networks

Amir Globerson

Our main goal is to build multivariate distributions that are compact and have guaranteed independence (or conditional independence; CI) properties. In this lecture, we will describe a class of such models called Bayesian Networks (BN).[1] A BN is defined using a directed graph and the properties of the directed graph imply the conditional independence properties of the BN.

1 Definitions and Notations

We begin with a few definitions and results that will be needed later.

1.1 Directed Graphs

Definition 1. *A* **directed graph** *is a pair (V, E) where V is a set of nodes (which we will typically denote by $\{1, \ldots, n\}$ and ordered pairs of vertices $(i, j) \in E$ (where (i, j) corresponds graphically to an edge $i \to j$). A Directed Acyclic Graph (DAG) is a directed graph that has no directed cycles.*

Definition 2. *A* **topological ordering** *of a DAG is an ordering of its nodes $i_1, \ldots, i_n$ where i_k appears before all the nodes j such that $(i_k, j) \in E$. In other words all the children of i_k appear after it in the ordering. It's easy to see that each DAG has such an ordering and it can be found efficiently. Note that it is not necessarily unique.*

Definition 3. *Given a node i we shall be specifically interested in several sets of nodes. The parents of i are all the nodes with edges going into i. We denote it by $Pa(i)$. The descendants of i are all the nodes with a directed path* from *i. The ancestors of i are all the nodes with a directed path* to *i. The non-descendants of i are all the nodes that are not descendants of i. We denote those by $ND(i)$.*

In what follows we will use graphs to define distributions on variables $x_1, \ldots, x_n$. We will identify variables with nodes in the graph. Thus, the set of variables that corresponds to the parents of i will be denoted by $X_{Pa(i)}$.

1.2 Conditional Independence

Independence is one of the most important properties of a distribution. It tells us something very important about the relations between variables. For example if we know that X_i is independent of X_j it means that setting the value of X_i will not change the distribution of X_j. However, in many cases variables are not independent but rather independent given other variables.

Definition 4. *Given three random variables X, Y, Z we say that X is conditionally independent of Y given Z if for all x, y, z values it holds that:*

$$p(x, y|z) = p(x|z)p(y|z) \tag{1}$$

[1] Introduced by Judea Pearl in his 1988 book: "Probabilistic Reasoning in Intelligent Systems"

The above can be seen to be equivalent to the condition $p(x|y,z) = p(x|z)$.[2] *We denote this relation by $X \perp Y | Z$.*

Conditional independence has several useful properties. We mention two below and will use others during the course.

Lemma 1. *The Decomposition lemma for CI: Let W, X, Y, Z be random variables. Then $X \perp Z, W | Y$ implies $X \perp Z | Y$. The proof is straightforward.*

Another almost trivial result is:

Lemma 2. *If $X \perp Y | Z$ then $X \perp Y, Z | Z$. This follows since:*

$$p(x|y,z,z) = p(x|y,z) = p(x|z) \tag{2}$$

Where the last equality follows from $X \perp Y | Z$.

2 Definition of Bayesian Network

A Bayesian Network is a model of a distribution $p(x_1, \ldots, x_n)$ that is constructed via a DAG. It has some very nice provable conditional independence properties as we show later.

Assume we are given a DAG G and a set of distributions $p(x_i | x_{Pa(i)})$. The Bayesian network $B = (G, p)$ is a distribution over $x_1, \ldots, x_n$ defined as:

$$p_B(x_1, \ldots, x_n) = \prod_i p(x_i | x_{Pa(i)}) \tag{3}$$

First, we need to prove it is indeed a distribution. Clearly it is non-negative. We also want to show that

$$\sum_{x_1, \ldots, x_n} p_B(x_1, \ldots, x_n) = 1 \tag{4}$$

To show this assume wlog that $1, \ldots, n$ is a topological ordering. Then:

$$p_B(x_1, \ldots, x_n) = p(x_1 | x_{Pa(1)}) p(x_2 | x_{Pa(2)}) \ldots p(x_n | x_{Pa(n)}) \tag{5}$$

where x_n appears only in the last term. Thus:

$$\begin{aligned} \sum_{x_1, \ldots, x_n} p_B(x_1, \ldots, x_n) &= \sum_{x_1, \ldots, x_{n-1}} p(x_1 | x_{Pa(1)}) p(x_2 | x_{Pa(2)}) \ldots \sum_{x_n} p(x_n | x_{Pa(n)}) \\ &= \sum_{x_1, \ldots, x_{n-1}} p(x_1 | x_{Pa(1)}) p(x_2 | x_{Pa(2)}) \ldots p(x_{n-1} | x_{Pa(n-1)}) \end{aligned}$$

Continuing this recursively we have that the overall sum is one.

3 Local Independence Properties of BNs

Next, we can ask what CI properties such a distribution may have. In other words if we know that p_B is a Bayesian Network for a graph G, what CI properties must it satisfy (regardless of the choice of p).

We need the following definition: the set of non-descendants of i are all the nodes that are not its children. Denote it by $ND(i)$.

[2] Some care should be taken for the case where $p(y|z) = 0$ to show the equivalence.

Theorem 1. *Given a Bayes net (G, p) the distribution p_B has the following properties:*

- *For all i, the variable X_i is conditionally independent of its non-descendants that are not his parents, given its parents:*
$$X_i \perp X_{ND(i)\setminus Pa(i)} | X_{Pa(i)} \tag{6}$$
- *It holds that:*
$$p_B(x_i|x_{Pa(i)}) = p(x_i|x_{Pa(i)}) \tag{7}$$

We note that from Lemma **??** *it also follows that: $X_i \perp X_{ND(i)} | X_{Pa(i)}$. This is the common form in the literature but can be confusing.*

Proof. First, consider the distribution $p_B(x_i, x_{ND(i)})$ which can easily be shown to have this form (after summing over all descendants of x_i)

$$p_B(x_i, x_{ND(i)}) = p(x_i|x_{Pa(i)}) \prod_k p(x_k|x_{Pa(k)}) \tag{8}$$

So that:

$$p_B(x_i|x_{ND(i)}) = \frac{p_B(x_i, x_{ND(i)})}{\sum_{x_i} p_B(x_i, x_{ND(i)})} = p(x_i|x_{Pa(i)}) \tag{9}$$

Or in other words:

$$p_B(x_i|x_{ND(i)\setminus Pa(i)}, x_{Pa(i)}) = p(x_i|x_{Pa(i)}) \tag{10}$$

Multiply by $p_B(x_{ND(i)\setminus Pa(i)}|x_{Pa(i)})$ to get:

$$p_B(x_i|x_{ND(i)\setminus Pa(i)}, x_{Pa(i)})p_B(x_{ND(i)\setminus Pa(i)}|x_{Pa(i)}) = p(x_i|x_{Pa(i)})p_B(x_{ND(i)\setminus Pa(i)}|x_{Pa(i)}) \tag{11}$$

Equivalent to:

$$p_B(x_i, x_{ND(i)\setminus Pa(i)}|x_{Pa(i)}) = p(x_i|x_{Pa(i)})p_B(x_{ND(i)\setminus Pa(i)}|x_{Pa(i)}) \tag{12}$$

Sum over all value of $x_{ND(i)\setminus Pa(i)}$ to get:

$$p_B(x_i|x_{Pa(i)}) = p(x_i|x_{Pa(i)}) \tag{13}$$

Substitute this in Eq. **??** to get:

$$p_B(x_i|x_{ND(i)\setminus Pa(i)}, x_{Pa(i)}) = p_B(x_i|x_{Pa(i)}) \tag{14}$$

From which it follows that $X_i \perp X_{ND(i)\setminus Pa(i)} | X_{Pa(i)}$.

To obtain $X_i \perp X_{ND(i)} | X_{Pa(i)}$ we just need to note that $X \perp Y | Z$ implies $X \perp Y, Z | Z$. □

The above theorem implies several things. First it means that p_B satisfies the following:

$$p_B(x_1, \ldots, x_n) = \prod_i p_B(x_i|x_{Pa(i)}) \tag{15}$$

We say that a distribution satisfying the above **factors according to** G.

Definition 1. *The Local Markov CI properties of DAG G are a set of n conditional independence properties given by:*

$$I_{LM}(G) = \{(X_i \perp X_{ND(i)} | X_{Pa(i)})\}_{i=1}^n \tag{16}$$

Definition 2. *Let p be a distribution over $X_1, \ldots, X_n$. We denote by $I(p)$ the set of* all *conditional independence properties that are true for p*

For example $I(p)$ can look something like:

$$I(p) = \{X_1, X_2 \perp X_3, X_4 | X_5, X_7, X_2, X_4 \perp X_5 | X_6\} \tag{17}$$

Using the above definitions Theorem **??** can be expressed as the following assertion: If p factorizes according to G then $I_{LM}(G) \subseteq I(p)$.

4 Local Markov leads to factorization

The previous section showed that any distribution that factorizes has the $I_{LM}(G)$ properties. Now, suppose that we would like to construct a distribution that has $I_{LM}(G)$ properties. We know we can use a Bayesian Network to obtain a distribution with these properties. But maybe there are other, more complex distributions that do not factor and have this property. The following theorem states that there are not (it is essentially the converse of Theorem **??**).

Theorem 2. *If p is a distribution such that $I_{LM}(G) \subseteq I(p)$, then p factorizes according to G.*

Proof. Assume wlog that $1, \dots, n$ is a topological ordering. Start by writing the chain rule:

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i | x_1, \dots, x_{i-1}) \tag{18}$$

We will be able to show the result if we show that $p(x_i|x_1, \dots, x_{i-1}) = p(x_i|x_{Pa(i)})$.

Because of the topological ordering, all of the parents of x_i are in $x_1, \dots, x_{i-1}$. Also, none of its descendants can be in this set. So we can write:

$$1, \dots, i-1 = Pa(i) \cup S \tag{19}$$

where $S \subseteq ND(i) \setminus Pa(i)$. Thus what we want to show can be rephrased as $p(x_i|x_{Pa(i)}, x_S) = p(x_i|x_{Pa(i)})$. In other words, we want to show that $X_i \perp X_S | X_{Pa(i)}$.

We know from the LM property that:

$$X_i \perp X_{ND(i) \setminus Pa(i)} | X_{ND(i)} \tag{20}$$

This is almost what we need, but remember that $S \subseteq ND(i)$. Fortunately, we have the Decomposition Lemma for CI (see above). Thus we can rewrite Eq. **??** as:

$$X_i \perp X_S, X_{ND(i) \setminus \{Pa(i), S\}} | X_{Pa(i)} \tag{21}$$

And from the lemma we conclude:

$$X_i \perp X_S | X_{Pa(i)} \tag{22}$$

which is what we needed. □

5 What other conditional independencies hold for Bayesian Networks?

The LM properties only tell us about n specific CIs. What about any other property $X \perp Y | Z$. How can we tell if it holds in a given BN or not? We may suspect that $X \perp Y | Z$ in a BN based on a graph if Z somehow blocks the influence of X on Y. The LM is an instance of this. We may thus first hypothesize that if there is no path between X and Y in the graph that does not go through Z the property will hold. This in fact is not true since the explaining away phenomenon is a counter example.

We need the following definition:

Definition 5. *An undirected path in a DAG G is active given a node set E if:*

- *For every V structure $i \rightarrow j \leftarrow k$ in the path either j or one of its descendants is in E.*
- *Every other node in the path is not in E*

The idea behind an active path is that if there is an active path between X_i and X_j given E then X_i is potentially dependent on X_j given E. This is extended to larger variable sets via the following definition.

Definition 6. *Given variables sets X, Y, Z in the graph G we say that X is d-separated from Y given Z if there is no active trail between a node in X and a node in Y given Z.*

The set of CI properties $I_{d-sep}(G)$ corresponds to all triplets $X \perp Y | Z$ where X is d-separated from Y given Z. In other words $X \perp Y | Z$ is in $d - sep(G)$ if every path between X and Y either has a node in Z that is not on a v-structure or it has a node outside Z that is in a v-structure or a descendant of a node in a v structure.

What can we say about d-separation and Bayesian networks?

Proposition 1. *For any BN on graph G it holds that $I(p) \supseteq I_{d-sep}(G)$.*

We will not show this here. It will wait until we talk about undirected models.

Since $I_{d-sep}(G)$ is quite a large set, we might expect that it captures all CI properties in a BN on G. This is clearly wrong however since for every graph G we can define a BN such that all variables are independent (i.e., $p_B = \prod_i p(x_i)$). Thus $I(p)$ will clearly contain more properties than those specified in $I_{d-sep}(G)$. So clearly $I_{d-sep}(G) \not\supseteq I(p)$.

However, you may suspect that this example is very specific. In other words we had to choose very specific p distributions to construct this p_B such that it will have properties not in $I_{d-sep}(G)$. In fact, it can be shown that except for a set of measure zero all BN on G will have satisfy $I(p) = I_{d-sep}(G)$.

In summary the following properties are equivalent:

- The distribution p factorizes according to G (i.e., it is a BN for G).
- $I(p) \supseteq I_{LM}(G)$
- $I(p) \supseteq I_{d-sep}(G)$