The Basics of Graphical Models

David M. Blei
Columbia University

October 3, 2015

Introduction

¶ These notes follow Chapter 2 of *An Introduction to Probabilistic Graphical Models* by Michael Jordan. Many figures are taken from this chapter.

¶ Consider a set of random variables $\{X_1, \ldots, X_n\}$. We are interested in

- Which variables are independent?
- Which variables are conditionally independent given others?
- What are the marginal distributions of subsets of variables?

These questions are answered with the *joint distribution* $P(X_1, \ldots, X_n)$.

- Marginalization is answered by summing over the joint.
- Independence is answered by checking factorizations.

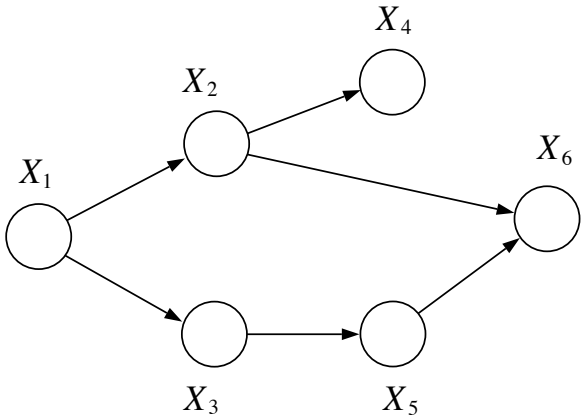
¶ Assume the variables are discrete (i.e., categorical). The joint distribution is a table $p(x_1, \ldots, x_n)$. Each cell is non-negative and the table sums to one. If there are r possible values for each variable then the naïve representation of this table contains r^n elements. When n is large, this is expensive to store and to use.

Graphical models provide a more economic representation of the joint distribution by taking advantage of local relationships between random variables.

Directed graphical models

¶ A directed graphical model is a directed acyclic graph. The vertices are random variables $X_1, \ldots, X_n$; edges denote the "parent of" relationship, where π_i are the parents of X_i.

Here is an example:



The random variables are $\{X_1, \ldots, X_6\}$, e.g., $\pi_6 = \{X_5, X_2\}$.

¶ The graph defines a factorization of the joint distribution in terms of the conditional distributions $p(x_i \mid x_{\pi_i})$.

$$p(x_{1:6}) \triangleq p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)p(x_6 \mid x_2, x_5)$$

In general,

$$p(x_{1:n}) \triangleq \prod_{i=1}^{n} p(x_i \mid x_{\pi_i}). \tag{1}$$

(Note that we can use a set in the subscript.) This joint is defined in terms of *local probability tables*. Each table contains the conditional probabilities of a variable for each value of the conditioning set.

¶ By filling in the specific values for the conditional distributions, we produce a specific joint distribution of the ensemble of random variables. Holding the graph fixed, we can change the local probability tables to obtain a different joint.

Now consider all possible local probability tables. We see that *the graphical model represents a family of distributions*. The family is defined by those whose joint can be written in terms of the factorization implied by the graph. It is important to notice that this is not all distributions over the collection of random variables.

Graphical models represent a family of distributions.

¶ What is the advantage of limiting the family? Suppose $x_{1:n}$ are binary random variables. The full joint requires 2^n values, one per entry. The graphical model joint requires $\sum_{i=1}^{n} 2^{|\pi_i|}$ entries. *We have replaced exponential growth in n by exponential growth in* $|\pi_i|$.

In statistical and machine learning applications, we represent data as random variables and analyze data via their joint distribution. We enjoy big savings when each data point only depends on a couple of parents.

¶ This is only part of the story. In addition to economic representation:

- Graphical models give us inferential machinery for computing probabilistic quantities and answering questions about the joint, i.e., the graph. The graph determines, and thus lets us control, the cost of computation. (And, as an aside, these same considerations apply when thinking about data and statistical efficiency. But this is less looked at in the graphical models literature.)
- Finally, graphs are more generic than specific joint distributions. A graphical model of binary data can be treated with similar algorithms as a graphical model with r-ary data. And, later, we will see how the same algorithms can treat discrete / categorical variables similarly to continuous variables, two domains that were largely considered separately. For example, graphical models connect the algorithms used to work with hidden Markov models to those used to work with the Kalman filter.

Directed graphical models and conditional independence

¶ Recall the definition of independence

$$x_A \perp\!\!\!\perp x_B \rightarrow p(x_A, x_B) = p(x_A)p(x_B) \tag{2}$$
$$\rightarrow p(x_A \mid x_B) = p(x_A) \tag{3}$$
$$\rightarrow p(x_B \mid x_A) = p(x_B) \tag{4}$$

And recall the equivalent definitions of conditional independence

$$x_A \perp\!\!\!\perp x_B \mid x_C \rightarrow p(x_A, x_B \mid x_C) = p(x_A, x_B \mid x_C) \tag{5}$$
$$\rightarrow p(x_A \mid x_B, x_C) = p(x_A \mid x_C) \tag{6}$$
$$\rightarrow p(x_B \mid x_A, x_C) = p(x_B \mid x_C) \tag{7}$$

These are questions about factorizations of marginal distributions. They can be answered by examining—or computing about—the graph.

¶ Recall the chain rule of probability

$$p(x_{1:n}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1}) \tag{8}$$

In our example

$$p(x_{1:6}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_6 \mid x_{1:5}). \tag{9}$$

The joint distribution defined by the graph is suggestive that, e.g.,

$$p(x_4 \mid x_1, x_2, x_3) = p(x_4 \mid x_2), \tag{10}$$

which means that

$$x_4 \perp\!\!\!\perp x_{\{1,3\}} \mid x_2. \tag{11}$$

This statement is true. It is one of the *basic conditional independence statements*.

Let's prove it:

- Write the conditional as a ratio of marginals

$$p(x_4 \mid x_1, x_2, x_3) = \frac{p(x_1, x_2, x_3, x_4)}{p(x_1, x_2, x_3)} \tag{12}$$

- Numerator: take the joint and marginalize out x_5 and x_6
- Denominator: Further marginalize out x_4 from the result of the previous step.
- Finally, divide to show that this equals $p(x_4 \mid x_2)$.

¶ More generally, let $\mathcal{I}$ be a *topological ordering* of the random variables, which ensures that π_i occurs in the ordering before i. Let v_i be the set of indices that appear before i, not including π_i. The set of basic conditional independence statements is

$$\{x_i \perp\!\!\!\perp x_{v_i} \mid x_{\pi_i}\} \tag{13}$$

In our example, one valid topological ordering is $\mathcal{I} = \{1, 2, 3, 4, 5, 6\}$. This implies the following independencies,

$$X_1 \perp\!\!\!\perp 0 \mid 0 \tag{14}$$
$$X_2 \perp\!\!\!\perp 0 \mid X_1 \tag{15}$$
$$X_3 \perp\!\!\!\perp X_2 \mid X_1 \tag{16}$$
$$X_4 \perp\!\!\!\perp \{X_1, X_3\} \mid X_2 \tag{17}$$
$$X_5 \perp\!\!\!\perp \{X_1, X_2, X_4\} \mid X_3 \tag{18}$$
$$X_6 \perp\!\!\!\perp \{X_1, X_3, X_4\} \mid \{X_2, X_5\} \tag{19}$$

¶ This is a little inelegant because it depends on an ordering. Here is a graph-specific definition of the basic conditional independencies. Redefine v_i to be all the ancestors of i excluding the parents. Using this definition, we can define the basic conditional independencies as in Equation (13).

Note: This does not give us all of the possible basic conditional independencies, i.e., all of those possible by traversing all topological orderings, but that doesn't really matter.) The point is this. *We can read off conditional independencies by looking at the graph.*

¶ We emphasize that these independencies hold regardless of the specific local probability tables. This gives us an insight about the nature of graphical models:

By using the cheaper factorized representation of the joint, we are making certain independence assumptions about our random variables.

This makes more precise—a little, anyway—the difference between the family specified by the graphical model and the family of all joints.

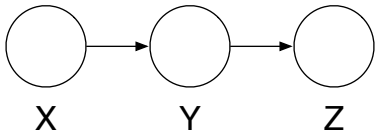
Q: Are these the only independence assumptions we are making?

The Bayes ball algorithm and d-separation

¶ Note that a node's parents separate it from its ancestors. It appears that conditional independencies are related to the graph and, in particular, to graph separation. We will next uncover the relationship between graph separation and conditional independence.

To do this, and deepen our understanding of independence and graphical models, we look at three simple graphs. We ask which independencies hold in these graphs and consider the relationship to classical graph separation.

¶ The first is a little sequence,



$$p(x, y, z) = p(x)p(y \mid x)p(z \mid y). \tag{20}$$

Here,

$$X \perp\!\!\!\perp Z \mid Y. \tag{21}$$

To see this,

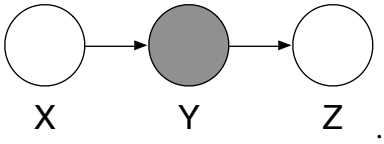
$$\begin{aligned} p(x \mid y, z) &= \frac{p(x, y, z)}{p(y, z)} && (22)\\ &= \frac{p(x)p(y \mid x)p(z \mid y)}{p(z \mid y)\sum_{x'} p(x')p(y \mid x')} && (23)\\ &= \frac{p(x, y)}{p(y)} && (24)\\ &= p(x \mid y) && (25) \end{aligned}$$

We assert that no other independencies hold. (E.g., is it *not* true that $X \perp\!\!\!\perp Z$.)

Here we do not need to do this algebra. This is one of the basic conditional independencies. The parent of Z is Y, and X is a non-parent ancestor of Z.

Important subtlety: This means that other independencies do not *necessarily* hold. For some settings of $p(y \mid x)$ it may be true that $X \perp\!\!\!\perp Z$. But, not for all. In other words, a more restrictive family of joints will be contained in the less restrictive family.

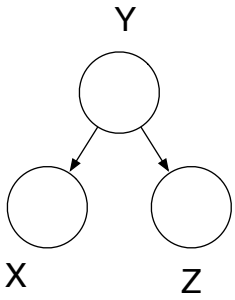
In this graph, conditional independence can be interpreted as graph separation. In graphical models notation, we shade the node that we are conditioning on:



We can see that Y separates X and Z.

The intuition: X is the "past", Y is the "present", Z is the "future". Given the present, the past is independent of the future. This is the *Markov assumption*. This graph is a three step *Markov chain*.

¶ The second graph is a little tree,



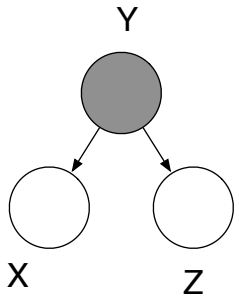
$$p(x, y, z) = p(y)p(x \mid y)p(z \mid y) \tag{26}$$

Here we have again that $X \perp\!\!\!\perp Z \mid Y$. We calculate that the conditional joint factorizes,

$$p(x, z \mid y) = \frac{p(y)p(x \mid y)p(z \mid y)}{p(y)} \tag{27}$$

$$= p(x \mid y)p(z \mid y) \tag{28}$$

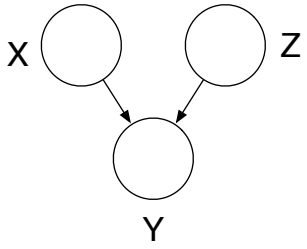
We assert that no other conditional independencies hold. Again, simple graph separation indicates independence,



The intuition behind this graph comes from a latent variable model. In our previous lecture, this graph describes the unknown coin flipped twice.

As another example, let X be "shoe size" and Z be "amount of gray hair". In general, these are dependent variables. But suppose Y is "age". Conditioned on Y, X and Z become independent. Graphically, we can see this. It is through "age" that "shoe size" and "gray hair" depend on each other.

¶ The last simple graph is an "inverse tree"



$$p(x, y, z) = p(x)p(z)p(y \mid x, z) \tag{29}$$

Here the only independence statement is $X \perp\!\!\!\perp Z$. In particular, it is *not* necessarily true that $X \perp\!\!\!\perp Z \mid Y$.

For intuition, think of a causal model: Y is "I'm late for lunch"; X is "I'm abducted by aliens", a possible cause of being late; Z is "My watch is broken", another possible cause. Marginally, being abducted and breaking my watch are independent. But conditioned on my lateness, knowing about one tells us about the likelihood of the other. (E.g., if I'm late and you know that my watch is broken, then this decreases the chance that I was abducted.)

Alas this independency does *not* correspond to graph separation.

¶ With these simple graphs in hand, we can now discuss d-separation, a notion of graph separability that lets us determine the validity of any conditional independence statement in a directed graphical model.

Suppose we are testing a conditional independence statement,

$$X_A \perp\!\!\!\perp X_B \mid X_C. \tag{30}$$

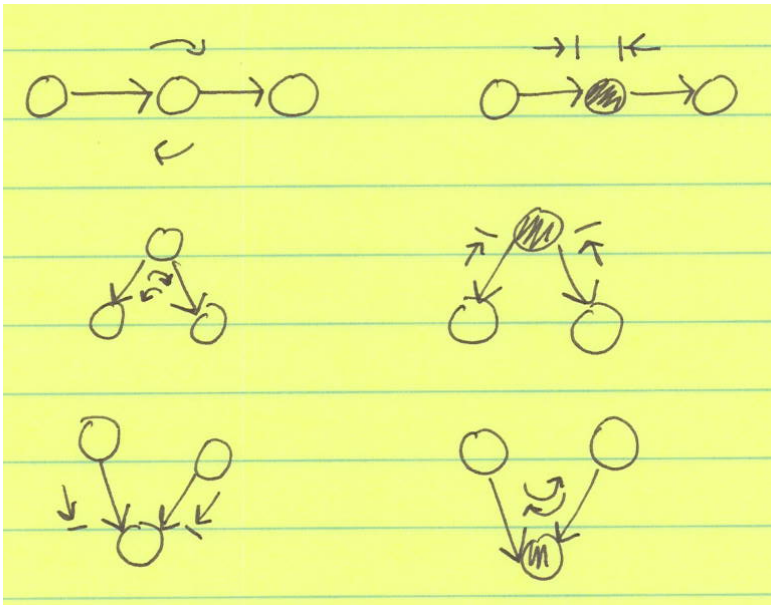
We shade the nodes being conditioned on. We then decide, using the "Bayes ball" algorithm, whether the conditioned nodes d-separate the nodes on either side of the independence relation.

The Bayes ball algorithm is a reachability algorithm. We start balls off at one of the sets of variables. If they can reach one of the other set then the conditional independence statement is false.

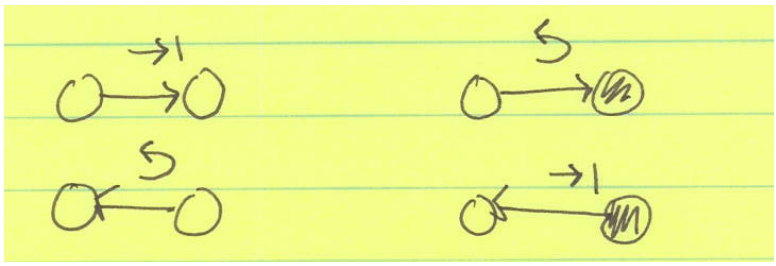
The balls bounce around the graph according to rules based on the three simple graphs. We consider a ball starting at X and going through Y on its way to Z. (To be clear, if the move is allowed, then the next step is for the ball to be at Y and we ask if it can go through Z en route to another node.)

Note that it does not matter if the source node X and destination node Z are shaded.

¶ Here are the rules:



In addition, there are rules derived by contemplating a ball going through a node and then back to the source node:



¶ Some examples:

1. Look at our example graph.

(a) $X_1 \perp\!\!\!\perp X_6 \,|\, \{X_2, X_3\}$? Yes.
(b) $X_2 \perp\!\!\!\perp X_3 \,|\, \{X_1, X_6\}$? No.

2. A Markov chain is the simple sequence graph with any length sequence. The basic conditional independencies are that

$$X_{i+1} \perp\!\!\!\perp X_{1:(i-1)} \,|\, X_i. \tag{31}$$

But Bayes ball tells us more, e.g.

$$\begin{aligned} X_1 \perp\!\!\!\perp X_5 \quad &| \quad X_4 \\ X_1 \perp\!\!\!\perp X_5 \quad &| \quad X_2 \\ X_1 \perp\!\!\!\perp X_5 \quad &| \quad X_2, X_4 \end{aligned}$$

3. Now consider a hidden Markov model which is used, for example, in speech recognition. The Bayes ball algorithm reveals that there are no conditional independencies among the observations.
4. (Optional) Look at a tree model, e.g., of genetic sequences in a family tree. What kinds of independencies do we see?
5. Look at a Bayesian hierarchical regression model. (E.g., consider testing in different schools.) How are the groups related? What if we know the prior?

¶ Remarks on Bayes ball:

- It's not an algorithm that is necessarily very interesting to implement. But it's very useful to look at graphs—i.e., at structured joint distributions—and understand the complete set of conditional independence and independence assumptions that are being made. As we have shown, this is not obvious either from the joint distribution or the structure alone.
- The idea of a ball bouncing around is a theme that we will come back to. It won't be balls, but be "messages" (i.e., information). Just as balls bouncing around the graph help us understand independence, messages traveling on the graph will help us make probabilistic computations.

The Hammersley-Clifford theorem

¶ Punchline:

Consider two families of joint probability distributions, both obtained from the graphical model G.

1. Family of joints found by ranging over all conditional probability tables associated with G.
2. All joints that respect *all* conditional independence statements, implied by G and d-separation.

The Hammersley-Clifford theorem says that these families are the same.

¶ More verbose:

We find the first family by varying the local conditional probability tables and computing the resulting joint from its factorization. This is what we meant earlier when we said that a graphical model defines a family of probability distributions.

We obtain the second family as follows. First, compute every conditional independence statement that is implied by the graph. (Use Bayes ball.) Then, consider *every* joint distribution of the same set of variables. Note this does not reference the local conditional probability tables. For each joint, check whether *all* the conditional independence statements hold. If one does not, throw the joint away. Those that remain are the second family.

The Hammersley-Clifford theorem says that these two families of joints—one obtained by checking conditional independencies and the other obtained by varying local probability tables—are the same.

As stated in the chapter, this theorem is at the core of the graphical models formalism. It makes precise and clear what limitations (or assumptions) we place on the family of joint distributions when we specify a graphical model.

Undirected graphical models

In this class, we will mainly focus on directed graphical models. However, undirected graphical models, which are also known as *Markov random fields*, are a very useful formalism as well. They are important to learn about to be fluent in graphical models, will be useful later when we talk about exact inference, and further refine the picture of the relationship between graphs and probability models.

A definition, via conditional independencies

When discussing directed models, we began with a definition of how to map a graph to a joint and then showed how the resulting conditional independencies can be seen from the graph. Here we will go in reverse. Based on an undirected graph, we will first define the conditional independencies that we want to hold in its corresponding joint distribution. We will then define the form of that distribution.

Consider an undirected graph $G = (V, E)$ and three sets of nodes A, B, and C. We will want a joint distribution such that $X_A \perp\!\!\!\perp X_C \mid X_B$ if X_B separates X_A and X_C, in the usual graph-theoretic sense of separate.

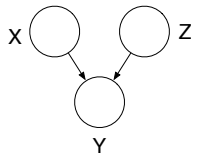
Formally, quoting from the book, "if every path from a node in X_A to a node in X_C includes at least one node in X_B then we assert that $X_A \perp\!\!\!\perp X_C \mid X_B$."

Again we emphasize that we are representing a *family* of distributions. These are the conditional independence statements that (we assert) have to hold. For various instantiations of the graphical model (i.e., various members of the family) other conditional independencies may also hold.

Undirected and directed graphical models are different

Consider the families of families expressable by directed and undirected graphical models, respectively. Not all directed graphical models can be written as an undirected graphical model, and vice versa.

First consider this directed graph,



As we said, the only conditional independence statement that is true for this graph is $X \perp\!\!\!\perp Z$. We cannot write an undirected graphical model such that this is the *only* conditional independence statement, i.e. where $X \not\perp\!\!\!\perp Z \mid Y$.

Now consider this undirected graph,

[Four nodes connected in a square]

This graph expresses the family characterized by the following independencies,

$$X \perp\!\!\!\perp Y \quad | \quad \{W, Z\} \tag{32}$$
$$W \perp\!\!\!\perp Z \quad | \quad \{X, Y\} \tag{33}$$

And these are the only ones.

We cannot write down a directed graphical model such that these are the only two conditional independence statements. Exercise: Confirm this.

Note that there are types of directed and undirected graphical models that can be written as either. We will one such important class when we talk about inference. But, in general, we have just demonstrated that they have different expressive power.

The joint distribution in an undirected graphical model

From the conditional independencies, we will now develop a representation of the joint distribution. Our goal is to represent the joint as a product of "local" functions, which we

will call *potential functions*, whose arguments are subsets of the random variables,

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{S \in \mathcal{S}} \psi(x_S) \tag{34}$$

Here S is a set of nodes, $\psi(\cdot)$ are arbitrary different potential functions (notation overloaded), and $\mathcal{S}$ is a collection of subsets. (We specify them later.) We would like these functions to be non-negative but otherwise arbitrary, so we will be satisfied with specifying them up to a scaling constant Z. (We use this notation to be consistent with the literature; Z is not a random variable.) This is called the *normalizing constant*, and will be an important quantity for much of this course.

We need to define what we mean by "local." This amounts to choosing the arguments of each of the potential functions, i.e., choosing the subsets $\mathcal{S}$. Let's return to the conditional independencies that we are hoping to assume with this representation. These imply that if two nodes X_1 and X_3 are separated by a third X_2 then $X_1 \perp\!\!\!\perp X_3 \mid X_2$. This implies that the conditional distribution factorizes,

$$p(x_1, x_3 \mid x_2) = p(x_1 \mid x_2) p(x_3 \mid x_2). \tag{35}$$

This further implies that the three nodes cannot participate in a single potential. Why? If there were an arbitrary potential function $\psi(x_1, x_2, x_3)$ in the joint distribution of Equation (34) then it would be impossible for the conditional (which, recall, is proportional to the joint) to factorize across x_1 and x_3.

Maximal cliques. This is suggestive that the potential functions should only be defined on *cliques*, which are sets of nodes that are fully connected, or subsets of cliques. Because of the direct connections between all nodes in a clique we are guaranteed that the graph does not imply any conditional independencies between them; thus it is safe to include them in the arguments to a potential. Conversely, as we argued above, if two nodes are not directly connected in the graph then there is a conditional independence statement that we can make. Thus, they should not appear together in a potential.

In the theory around undirected graphical models, the joint is defined on the set of *maximal cliques*, i.e., completely connected components of the graph that cannot be expanded without breaking complete connectedness. Every node is part of a maximal clique. Thus, we can write the joint as

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi(x_C). \tag{36}$$

Here, $\mathcal{C}$ is the set of maximal cliques and C is a particular clique (i.e., set of nodes). The normalizing constant is

$$Z = \sum_x \prod_{C \in \mathcal{C}} \psi(x_C). \tag{37}$$

It is difficult to compute. (Why?) We'll come back to that later in the semester.

This joint distribution respects the set of conditional independence statements implied by usual graph separability on the underlying graph.

Finally, in practice we often define undirected graphical models in terms of other cliques, in addition to or instead of maximal cliques. As long as we don't steer beyond a maximal clique, this preserves the relationship between graph separation and conditional independence.

Interpreting potentials. The potential functions we set up are arbitrary positive valued functions. They are *not* conditional probabilities (necessarily) as in the directed graphical models case. However, they can be interpreted as providing "agreement" to configurations of variables that have high probability. If the potential on $\psi(x_1, x_2)$ is high then the configuration with those values has higher probability. Though we will not discuss it in depth, this is how undirected graphical models play a large role in statistical physics (the field in which they were invented).

Hammersley-Clifford for undirected graphical models. We can state a similar theorem for undirected graphical models as we did for directed graphical models.

Fix an undirected graph G.

Define one family of distributions by ranging over all possible potential functions over the maximal cliques of the graph, and calculating the joint distribution in Equation (36).

Define a second family of distributions by looking at all joint distributions over the set of nodes in the graph and filtering out only those for which the set of conditional independence statements—defined by graph separability—holds.

These two sets of distributions are the same.