10-708: Probabilistic Graphical Models 10-708, Spring 2017

2 : Directed GMs: Bayesian Networks

Lecturer: Eric P. Xing *Scribes: Jayanth Koushik, Hiroaki Hayashi, Christian Perez*

Topic: Directed GMs

1 Types of Graphical Models

The goal of probabilistic graphical models is to establish a formal relation between the graph (edges, nodes) with its corresponding probability distribution. Although there are numbers of alternatives that also use a graph (e.g. deep neural network), it's not precise what edges mean in these alternatives. We distinguish two types of graphical models as follows:

1. **Directed GM**: Directed edges give causality relationships between random variables.
2. **Undirected GM**: Undirected edges give correlations between random variables.

2 Notations

To represent a GM, we specify certain notations for each important concept.

- **Variable** : X, a placeholder of realized values (x). E.g. $x \in [0, 1]$. Also, we can index: $X_i^{(i)}$. Typically, the lower case index is the dimension while the uppercase is the data index (i.e. number of observations).
- **Random variable** : X
- **Random vector**: $\boldsymbol{X}$
- **Random matrix**: $\boldsymbol{X} = \{X_{i,j}\}$
- **Parameters**: Typically use Greek letters, e.g. α, β, θ. They can also be indexed.

3 Example: A Dishonest Casino

Consider a die game with both a fair die and a loaded die. Suppose the dealer switches back and forth between the die once in a while. As a player, one may ask the following three questions given the observations (a number sequence X). Note that X is a partially observed data, since we have no information regarding which die was used throughout the game. We call this binary sequence of the choice of dice as Y.

1. **Evaluation** : The likelihood of the observed sequence, given the model in your mind. In other words, this is a problem to estimate $p(X)$.

2. **Decoding** : Underlying fact of Which parts of the sequence were generated by the loaded die. This problem's goal is to model $p(Y|X)$.
3. **Learning** : "Loadedness" or "Fairness" of each die. The goal of this problem is to discover the parameter(s) that characterize the problem. (e.g. $\theta^* = \text{argmax} f$ given some function f).

This game can be formalized as a simple graphical model as shown in Figure 8 (see Section 11).

Picking a graphical model for a given problem involves multiple considerations,: how to pick variables, structures, and probabilities. Depending on the problem, some variables are observed, some are hidden, they can be either continuous or discrete, etc. In terms of structure, one must consider designing a structure that realizes good balance between usefulness and simplicity. For example, *causal*:$y_1 \to x_1$, $y_2 \to x_2$.[1]

4 Bayesian Network

A Bayesian Network is a directed acyclic graph (DAG) whose nodes correspond to the random variables and edges correspond to the direct influence of the parent variable to the child variable. Bayesian Networks can represent the joint probability of the random variables in a factorized way, which achieves a compact representation for a set of conditional independence assumptions about the distribution. Each variable (node) is a stochastic function of its parents.

4.1 Factorization Theorem

Given a DAG, the most general form of the probability distribution that is **consistent** with the graph factors according to "**node given its parents**":

$$P(X) = \prod_{i=1:d} P(X_i|X_{\pi_i})$$

where X_{π_i} is the set of parent nodes of X_i, and d is the number of nodes. See Figure 1 for an example.

This graph can be factored and represented as follows:

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) =$$
$$P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2)P(X_6|X_3, X_4)P(X_7|X_6)P(X_8|X_5, X_6)$$

4.2 Local Structures and Independencies

Graphical models have three fundamental local structures that composes bigger graph structures. See Figure 2 for visualization.

- **Common parent** : Fixing B decouples A and C. When two random variables A and C have a common parent B, conditional independence $A \perp C|B$ holds.

[1] Here, the two causal statements are independent of each other

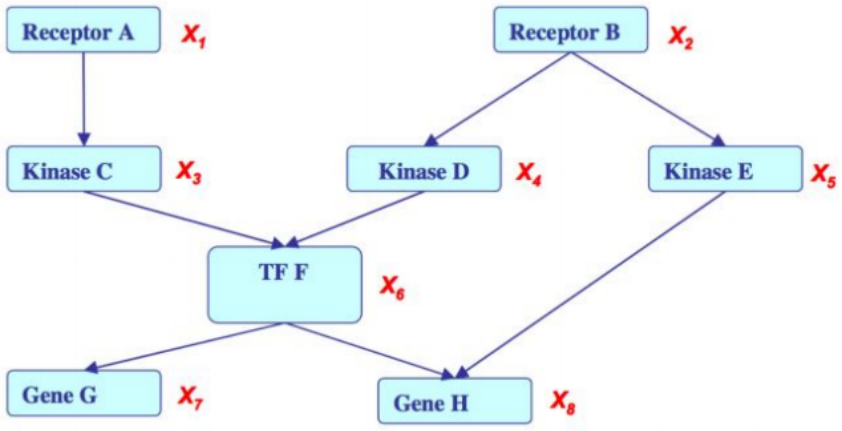
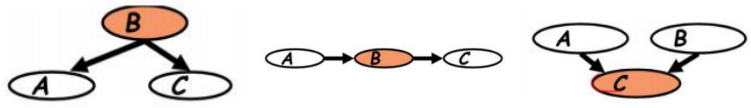


Figure 1: Example graph



(a) Common Parent (b) Cascade (c) V-structure

Figure 2: Different type of local structures.

- **Cascade** : Knowing B decouples A and C. When a middle node in a cascaded three random variables is known, a conditional independence $A \perp C|B$ holds.

- **V-structure** : If C is not observed, then A and B are independent. However, if it is given, then the independence is lost. (A and B are not independent given C). In this case, A and B are *marginally independent*.

The unintuitive V-structure can be described by a simple example. Suppose A = clock on tower, B = H bridge has high traffic, and C = Eric on time for class. If Eric is not on time and the clock is on time – then, our belief that B occurred is higher.

5 I-Maps

Definition : Let P be a distribution over X. I-map $I(P)$ is defined to be the set of independence assertions of the form $(X \perp\!\!\!\perp Y|Z)$ that hold in P.

Definition : Let K be an any graph object associated with a set of independencies $I(K)$. Then K is an *I-map* for a set of independencies I if $I(K) \subseteq I$.

For example, if a graph K is totally connected, then $I(K) = \emptyset \subset P$. A complete graph is "useless", since it does not give any knowledge about the structural knowledge.

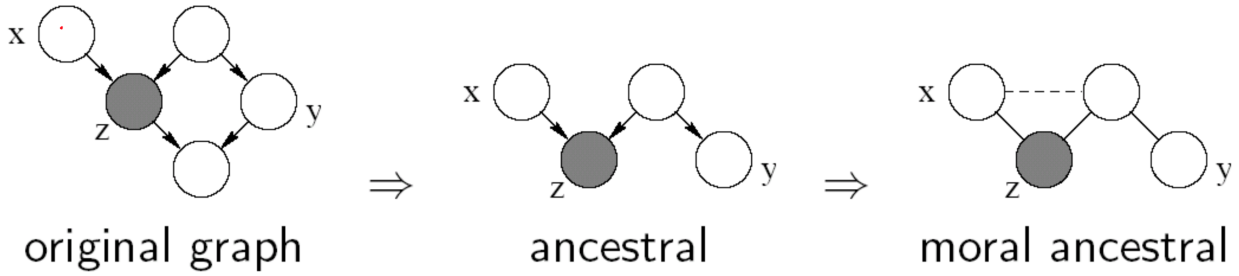


Figure 3: Converting a graph to a moralized ancestral graph for testing d-separation.

5.1 Facts about I-Map

For a graph G to be an I-map of P, any independence that G asserts must also hold in P. On the other hand, P may have additional independencies that are not included in G.

5.2 Local Conditional Independence Assumption

Given a Bayesian Network G, the following assertion holds.

Definition : Let Pa_{X_i} be the parents of X_i in G, and Nd_{X_i} be the variables that are *not* descendants of X_i. Then the following set of local conditional independence assumptions hold:

$$I_l(G) : \{X_i \perp\!\!\!\perp Nd_{X_i} | Pa_{X_i} : \forall i\}$$

In other words, a node X_i is independent of any non descendants given its parents.

6 D-separation

D-separation (D is for directed) can be defined in two ways.

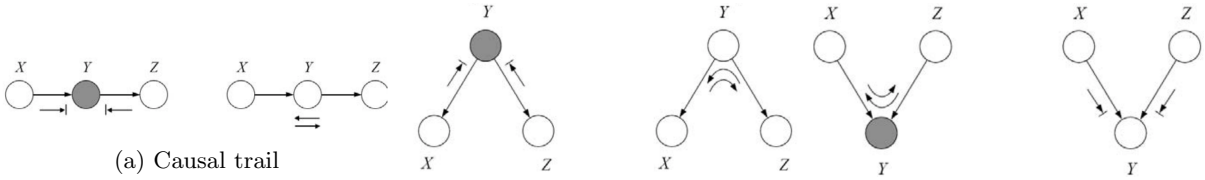
6.1 D-separation Through Moralized Ancestral Graph

Let X, Y, and Z be variables in a graph. We say X is d-separated from Y given Z if X and Y are separated in the moralized ancestral graph. This graph is constructed by first removing the descendants of X, Y, Z. Then we convert the edges to undirected edges and add edges between nodes which have a common child in the original graph, but no edge between them. This process is illustrated in Figure 3.

6.2 D-separation through Bayes Ball Algorithm

D-separation can also be defined through the Bayes ball algorithm. For this, we first consider these trails which are also illustrated in Figure 4.

- Causal trail $X \to Z \to Y$, and evidential trail $X \leftarrow Z \leftarrow Y$: active iff Z is not observed. This is shown in



(a) Causal trail (b) Common cause (c) Common effect

Figure 4: The different trails relevant to the Bayes ball algorithm, and the conditions under which they are active.

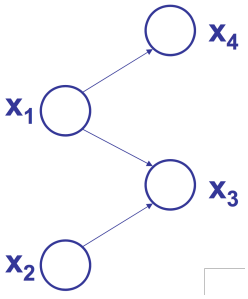


Figure 5: A directed graph

- Common cause $X \leftarrow Z \rightarrow Y$: active iff Z is not observed
- Common effect $X \rightarrow Z \leftarrow Y$: active iff Z or one of its descendants is observed

The Bayes ball algorithm works by placing balls at each node in X and moving them along active trails as described above. If no ball reaches a node in Y, we assert that X is d-separated from Y given Z.

D-separation can be used to find the global Markov properties of the graph. This is characterized by $I(G)$ which is a set of all independence properties that correspond to d-separation in a graph. For example, for the graph shown in Figure 5,

$$I(G) = \{X_2 \perp\!\!\!\perp \{X_1, X_4\},\ X_2 \perp\!\!\!\perp X_1 | X_4,\ X_2 \perp\!\!\!\perp X_4 | X_1,\ X_3 \perp\!\!\!\perp X_4 | X_1,\ X_3 \perp\!\!\!\perp X_4 | \{X_1, X_2\},\ X_2 \perp\!\!\!\perp X_4 | \{X_1, X_3\}\}$$

7 Equivalence theorem

Separation properties in a graph imply conditional independence in a distribution. This is captured by the equivalence theorem: For a graph G, let $\mathcal{D}_1$ be the set of distributions that satisfy $I(G)$. And let $\mathcal{D}_2$ be the set of distributions that factorize according to G i.e. $P(X) = \prod_{i=1}^{d} P(X_i | X_{\pi_i})$. Then $\mathcal{D}_1 = \mathcal{D}_2$. For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents. Based on the equivalence theorem, this can be done with conditional probability tables for discrete variables, or conditional density functions for continuous variables.

	b_0	b_1
a_0	0.4	0.6
a_1	0.4	0.6

Table 1: The distribution specified in this table factorizes according to the graph $A \to B$ but A is independent of B.

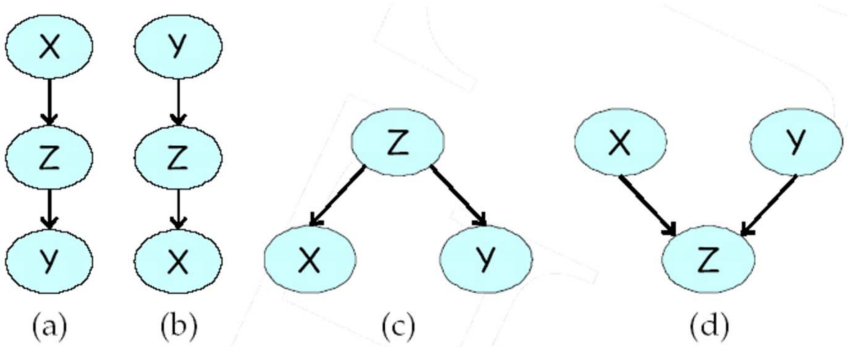


Figure 6: Four I-equivalent graphs.

8 Soundness and completeness

Soundness: If a distribution P factorizes according to a graph G, then $I(G) \subseteq I(P)$. **Completeness**: For any distribution P that factorizes over G, if $X \perp\!\!\!\perp Y | Z \in I(P)$, then over G, X and Y are d-separated given Z.

However, it is important to note that if X and Y are not d-separated given G, then it is not the case that X and Y are dependent given Z in *all* distributions that factorize over G. For example, consider the graph $A \to B$. Clearly, A and B are dependent. Note that every distribution over A and B factorizes according to this graph, since it is always true that $P(A, B) = P(A)P(B|A)$. But if we consider the specific distribution give in Table 1, then $A \perp\!\!\!\perp B$. However, we can assert that if X and Y are not d-separated given Z, then there is at least one distribution which factorizes according to the graph, and where X is not independent of Y given Z. Combining this with the above theorems gives us an important result.

For almost all distributions P that factorize over a graph G, i.e. except for a set of measure 0 in the space of CPD parameterizations, we have $I(P) = I(G)$.

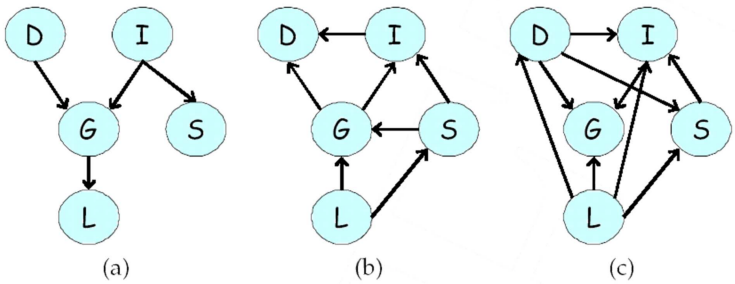


Figure 7: Equivalent minimal I-maps.

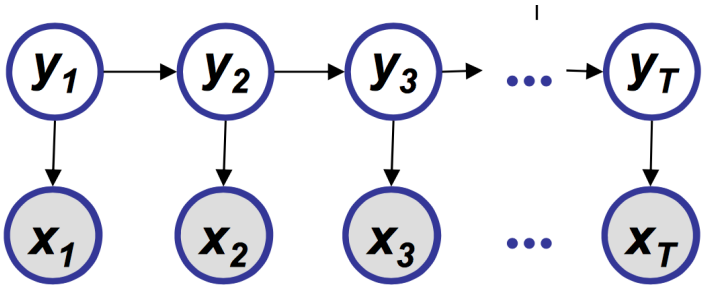


Figure 8: A Hidden Markov Model

9 Uniqueness of Bayesian Networks

Different directed graphs can be equivalent in the sense that they encode the same set of conditional independence properties. For example, all four graphs in Figure 6 have the same conditional independence $X \perp\!\!\!\perp Y|Z$. This notion can be captured with the concept of I-equivalence.

I-equivalence: Two graphs G_1 and G_2 are I-equivalent if $I(G_1) = I(G_2)$.

This is an equivalence relation over the set of all graphs. This has an important effect over determining the influence between variables. If $X \to Y$ in one graph, it is possible that $Y \to X$ in an I-equivalent graph.

I-equivalence can be determined by looking at the skeleton graphs, i.e. the graphs formed by converting directed edges to undirected edges. Two graphs are I-equivalent if they have the same skeleton, and the same set of v-structures.

10 Minimal I-map

It is trivial to find an I-map for a distribution since the complete graph is an I-map for *every* distribution. So it is useful to define the concept of a minimal I-map, which stops being an I-map on removing any single edge. However, the minimal I-map is also not unique as shown in Figure 7.

11 Example: Hidden Markov Models

A Hidden Markov Model (HMM) is a directed graphical model with hidden and observed variables. The hidden (latent) random variables y_t satisfy the Markov property i.e. y_t only depends on y_{t-1}. The observed variables x_t on the other hand, only depend on the corresponding hidden variables y_t. This is shown in Figure 8. More formally, the HMM is characterized by a set of transition probability distributions $P(y_t|y_{t-1})$, a start probability distribution $P(y_1)$, and a set of emission probability distributions $P(x_t|y_t)$. HMMs are used in applications such as speech and handwriting recognition.

12 Summary

- A Bayesian Network is a pair (G, P) where P factorizes over G, and where P is specified as a set of local conditional probability distributions over the nodes of G.

- Bayesian networks capture causality or generative schemes between entities.
- Local and global independence criteria can be identified from a graph using d-separation and the Bayes ball algorithm.
- Computing joint likelihoods is easy since it amounts to multiplying CPDs, but computing marginals, and thus inference in general, is hard.
- A causal scheme is not by itself Bayesian, so using directed graphical models does not necessarily imply a Bayesian approach.