

# Directed Graphical Models

Felipe José Bravo Márquez

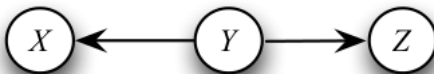
August 13, 2021

# Directed Graphical Models

- Probabilistic graphical models (PGMs) are a rich framework for encoding probability distributions over complex domains [Koller and Friedman, 2009].
- In this class we will focus on directed graphical models (DGMs), which are one type of PGM.
- Directed graphical models (DGMs) are a family of probability distributions that admit a compact parametrization that can be naturally described using a directed graph.
- DGMs are also known as Bayesian networks.
- Statistical inference for DGMs can be performed using frequentist or Bayesian methods, so it is misleading to call them Bayesian networks [Wasserman, 2013].

# Directed Acyclic Graphs (DAGs)

- A directed graph consists of a set of nodes with arrows between some nodes.
- Graphs are useful for representing independence relations between variables.
- More formally, a directed graph  $G$  consists of a set of vertices  $V$  and an edge set  $E$  of ordered pairs of vertices.
- For our purposes, each vertex corresponds to a random variable.
- If  $(Y, X) \in E$  then there is an arrow pointing from  $Y$  to  $X$ .



**Figure:** A directed graph with vertices  $V = \{X, Y, Z\}$  and edges  $E = \{(Y, X), (Y, Z)\}$ .

# Directed Acyclic Graphs (DAGs)

- If an arrow connects two variables  $X$  and  $Y$  (in either direction) we say that  $X$  and  $Y$  are adjacent.
- If there is an arrow from  $X$  to  $Y$  then  $X$  is a parent of  $Y$  and  $Y$  is a child of  $X$ .
- The set of all parents of  $X$  is denoted by  $\pi_X$  or  $\pi(X)$ .
- A directed path between two variables is a set of arrows all pointing in the same direction linking one variable to the other such as the chain shown below:

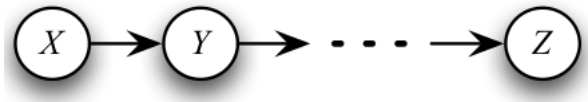


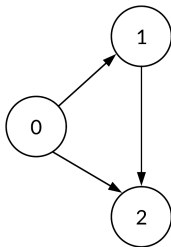
Figure: A chain graph with a directed path.

- $X$  is an ancestor of  $Y$  if there is a directed path from  $X$  to  $Y$  (or  $X = Y$ ).
- We also say that  $Y$  is a descendant of  $X$ .

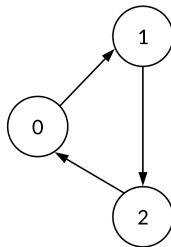
# Directed Acyclic Graphs (DAGs)

- A directed path that starts and ends at the same variable is called a cycle.
- A directed graph is acyclic if it has no cycles.
- In this case we say that the graph is a directed acyclic graph or DAG.

Acyclic Graph



Cyclic Graph



- From now on, we only deal with directed acyclic graphs since it is very difficult to provide a coherent probability semantics over graphs with directed cycles.
- For the remainder of this class we will use the terms Bayesian Network, directed graphical model (DGM) and directed acyclical graph (DAG) interchangeably.

# Probability and DAGs

- An important concept we need to introduce to understand DAGs is the chain rule of probability.
- For a set of random variables  $X_1, \dots, X_n$  we can write the joint probability function  $f(x_1, x_2, \dots, x_n)$  as

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

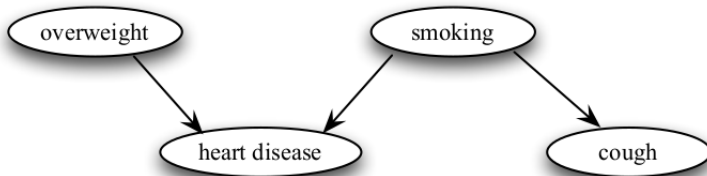
- A DAG is a distribution in which each factor on the right hand side depends only on a small number of ancestor variables  $\pi(x)$ . [?]
- Let  $G$  be a DAG with vertices  $V = (X_1, \dots, X_d)$ .
- If  $P$  is a distribution for  $V$  with probability function  $f(x)$  (density or mass), we say that  $G$  represents  $P$ , if

$$f(x) = \prod_{j=1}^d f(x_j|\pi_{x_j})$$

where  $\pi_{x_j}$  is the set of parent nodes of  $X_j$

# Probability and DAGs

- The next figure shows a DAG with four variables.



- The probability function takes the following decomposition:
- $f(\text{overweight}, \text{smoking}, \text{heart disease}, \text{cough}) = f(\text{overweight}) \times f(\text{smoking}) \times f(\text{heart, disease} | \text{overweight}, \text{smoking}) \times f(\text{cough} | \text{smoking})$ .

# Conditional Independence

- Let  $X$ ,  $Y$  and  $Z$  be random variables.
- $X$  and  $Y$  are conditionally independent given  $Z$ , written  $X \perp Y|Z$ , if:

$$f(x, y|z) = f(x|z)f(y|z)$$

for all  $x$ ,  $y$  and  $z$ .

- Notice that  $f$  can be either a density function for continuous random variables or a probability mass function for discrete random variables.
- Intuitively, this means that, once you know  $Z$ ,  $Y$  provides no extra information about  $X$ .



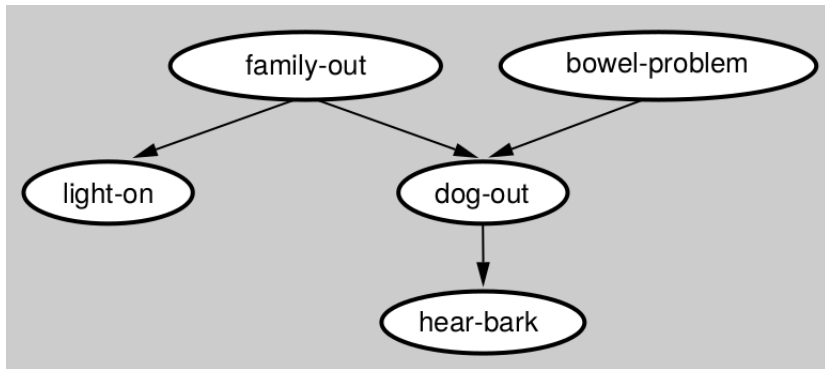
# An Example

- The best way to understand DAGs is to imagine trying to model a situation in which causality plays a role.
- And also our understanding of what is actually going on is incomplete
- So we need to describe things probabilistically.
- The following example is based on [?].
- Eugene Charniak is a famous AI researcher who's got the following situation.
- When he goes home at night, he wants to know if his family is home before trying the doors.
- Often when his wife leaves the house, she turns on an outdoor light.

# An Example

- Eugene's wife can also turn on the outdoor light if she is expecting a guest.
- Also, they have a female dog.
- When nobody is home, the dog is put in the back yard.
- The same is true if the dog has bowel troubles.
- Finally, if the dog is in the backyard, Eugene's will probably hear her barking
- The next slide shows a DAG encoding all the above causal relationships.

# An Example



- The DAG can help to predict what will happen in a particular scenario (if his family goes out, the dog goes out)
- Or to infer causes from observed effects (if the light is on and the dog is out, then his family is probably out).

- sdsad

- sdsad

# Estimation for DAGs

- Two estimation questions arise in the context of DAGs.
- First, given a DAG  $\mathcal{G}$  and data  $V_1, \dots, V_n$  from a distribution  $f$  consistent with  $\mathcal{G}$ , how do we estimate  $f$ ?
- Second, given data  $d_1, \dots, d_n$  how do we estimate  $\mathcal{G}$ ?
- The first question is pure estimation while the second involves model selection.
- These are very involved topics and are beyond the scope of this course.
- We will just briefly mention the main ideas.

# Estimation for DAGs

- If we are doing frequentist inference, we typically use some parametric model  $f(x|\pi_x; \theta_x)$  for each conditional density.
- The likelihood function is then

$$\mathcal{L}(\theta) = \prod_{i=1}^n f(d_i; \theta) = \prod_{i=1}^n \prod_{j=1}^m f(X_{ij}|\pi_j; \theta_j)$$

- where  $X_{ij}$  is the value of  $X_j$  for the  $i$ th data point and  $j$  are the parameters for the  $j$ th conditional density.
- We can then estimate the parameters by maximum likelihood.
- On the other hand, if we want to perform Bayesian inference we must set priors for all our variables  $X_1, \dots, X_m$  and estimate the posterior accordingly.

- To estimate the structure of the DAG itself, we could fit every possible DAG using maximum likelihood and use AIC (or some other method) to choose a DAG.
- However, there are many possible DAGs so we would need much data for such a method to be reliable.
- Also, searching through all possible DAGs is a serious computational challenge.
- Producing a valid, accurate confidence set for the DAG structure would require astronomical sample sizes.
- If prior information is available about part of the DAG structure, the computational and statistical problems are at least partly ameliorated [Wasserman, 2013].



# Conclusions

- Blabla

# References I



Koller, D. and Friedman, N. (2009).

*Probabilistic graphical models: principles and techniques.*  
MIT press.



Wasserman, L. (2013).

*All of statistics: a concise course in statistical inference.*  
Springer Science & Business Media.