Bayesian Networks — An Introduction

- Administrative: (i) please post for this week, (ii) this week largely technical set-up for last few weeks, (iii) paper topics (let's talk about these soon).
- A left-over from last time: Reduction & Circularity
  - "Identifiability" — the new buzzword
- Brief history of causal modelling with graphs and equations
  - Sewall Wright's (1920's, UW theoretical biologist) Path Analysis
  - Simon and Blalock (picked-up on Wright's theme, 50's–60's)
  - General Structural Equation Modeling (last 20-30 years)
- Bayesian Networks I — Evidential Networks
  - Conditional Independence — the central technical concept
  - DAGs as tools to represent Conditional Independence Structures
  - *d*-separation and conditional independence in DAGs (+ examples)

Issues of Reduction and Circularity

> In order to determine whether a probabilistic reduction of causation is possible, the central issue is not whether the word 'cause' appears in both the *analysandum* and the *analysans*; rather, the key question should be whether, given an assignment of probabilities to a set of factors, there is a unique set of causal relations among those factors compatible with the probability assignment and the theory in question. [Hitchcock's PC]

- Following Hitchcock, suppose that a set of factors, and a system of causal relations among those factors is given: call this the *causal structure* CS. And, let T be a theory connecting causal relations among factors with probabilistic relations among factors (*e.g.*, (*PRK*)).
- Then the causal structure CS will be *probabilistically identifiable relative to* T, if for every (this can be weakened) assignment of probabilities to the factors in CS that is compatible with CS and T, CS is the *unique* causal structure compatible with T and those probabilities.
- Intuitively, T allows you to infer that the causal structure is in fact CS, given the probability relations between factors.

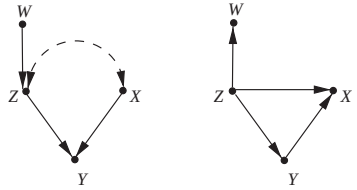
- Given a "reductive" probabilistic theory of causation T, there are many identifiability properties it might have, for instance:
  1. All causal structures are probabilistically identifiable relative to T.
  2. All causal structures having some interesting property are probabilistically identifiable relative to T.
  3. Any causal structure can be embedded in a causal structure that is probabilistically identifiable relative to T.
  4. The actual causal structure of the world (assuming there is such a thing) is probabilistically identifiable relative to T.
- Which (if any) of these is required for T to count as a successful "reduction" of causation to probability?
- One of the problems with traditional philosophical theories T of p-causation, is that it is unclear which causal structures are identifiable, relative to T. This is an advantage of the more recent "network" theories.
- As we will see, the recent "network" theories are not nearly as ambitious as the philosophical theories have been [they seem to aim for (2), (3), or (4), not (1)].

Some History of Statistical Causal Modeling

- In the 1920's the (UW) biologist Sewall Wright combined directed graphs and linear regression models to invent *path analysis* (these models were later more deeply studied, and generalized, by Herbert Simon and Hubert Blalock).
- In more recent years, many people have contributed to the more general field of *structural equation modeling*, which includes non-linear regression models.
- All of these models can be subsumed under a more general class of structures called *conditional independence structures* (see Wermuth & Lauritzen's survey).
- All of these models share two things: (1) they all use directed graphs as visual aides for representing statistical models, and (2) they all presuppose that certain *conditional independence relationships* are encoded (implicitly) by such graphs.
- I will begin by discussing directed acyclic graphs DAGs (the type of directed graph used in Bayesian networks). Then, I will move on to a discussion of conditional independencies and their representation in DAGs. Next time, I'll discuss how structural (regression) equations underlie the DAGs.

Directed Acyclic Graphs — DAGs

- A *graph* is a set of dots called *vertices* (nodes) connected by links called *edges*. If two vertices are connected by an edge, then they are said to be *adjacent*.
- *Directed* graphs (DGs) can have directed cycles ($X \leftrightarrow Y$, indicating a latent CC) but no self-loops ($X \nrightarrow X$). Directed acyclic graphs (DAGs) have *no* cycles.
- A *path* is an unbroken, nonintersecting route traced along the edges of a graph. For instance, $\{\langle W, Z\rangle, \langle Z, Y\rangle, \langle Y, X\rangle, \langle X, Z\rangle\}$ is a path in graph (a), below.



- A path is said to be *directed*, if every pair $\langle X, Y\rangle$ in the path is such that $X \rightarrow Y$.

For instance, the path $\{\langle W, Z\rangle, \langle Z, Y\rangle\}$ in (a), above, is directed, but $\{\langle W, Z\rangle, \langle Z, Y\rangle, \langle Y, X\rangle\}$ and $\{\langle W, Z\rangle, \langle Z, X\rangle\}$ are not (since $X \rightarrow Y$ and $Z \leftrightarrow X$).

- The *skeleton* of a graph is obtained by removing all arrows from the graph [for instance, the DAGs (a) and (b), above, have the same skeleton].
- If there is (is not) a path connecting X and Y, then X and Y are *(dis)connected*.
- We use kinship relations (*e.g.*, parents, children, ancestors, descendants, spouse) to describe vertices. In (a), Y has three ancestors (X, Z, W) and two parents (X, Z). And, X has no parents ($\therefore$ no ancestors), one spouse (Z) and one child (Y).
- A *family* in a graph is a set of nodes containing a node and all its parents. For instance, in (a), the only families are $\{W\}$, $\{X\}$, $\{Z, W\}$, and $\{Y, Z, X\}$.
- A *root* has no parents and a *sink* has no children. A connected DAG in which every node has at most one parent is called a *tree*. A tree in which every node has at most one child is a *chain*. If every pair is adjacent, the DAG is *complete*.
- So, (a) is connected but not complete, since $\{W, X\}$ and $\{W, Y\}$ are not adjacent.

Conditional Independence Between Sets of Variables

- Let $V = \{V_1, \ldots, V_n\}$ be an (partially) ordered set of random variables (say, in a DAG $\mathcal{G}$, where $\rightarrow$ imposes the order), and let X, Y, and Z be three subsets of V. And, let $\Pr(v) = \Pr(\bigwedge_i V_i = v_i)$ be the joint probability distribution over the V_i.
- We use the notation $(X \perp\!\!\!\perp Y \mid Z)$ to say that *X and Y are conditionally independent, given Z*. $(X \perp\!\!\!\perp Y \mid Z)$ iff $\Pr(x \mid y \,\&\, z) = \Pr(x \mid z)$ [$\Pr(y \,\&\, z) > 0$].
- That is, $\Pr(\bigwedge_i X_i = x_i \mid \bigwedge_j Y_j = y_j \,\&\, \bigwedge_k Z_k = z_k) = \Pr(\bigwedge_i X_i = x_i \mid \bigwedge_k Z_k = z_k)$, for *all* X_i, x_i, Y_j, y_j, Z_k, z_k. That is, all values of all members of X are screened-off from all values of all members of Y, by all values of all members of Z.
- Some important properties of the relation $(X \perp\!\!\!\perp Y \mid Z)$ are as follows:
  - Symmetry: $(X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (Y \perp\!\!\!\perp X \mid Z)$
  - Decomposition: $(X \perp\!\!\!\perp Y \cup W \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid Z)$
  - Weak Union: $(X \perp\!\!\!\perp Y \cup W \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid Z \cup W)$
  - Contraction: $(X \perp\!\!\!\perp Y \mid Z) \,\&\, (X \perp\!\!\!\perp W \mid Z \cup Y) \Longrightarrow (X \perp\!\!\!\perp Y \cup W \mid Z)$

  - Intersection: $(X \perp\!\!\!\perp W \mid Z \cup Y) \,\&\, (X \perp\!\!\!\perp Y \mid Z \cup W) \Longrightarrow (X \perp\!\!\!\perp Y \cup W \mid Z)$ (Intersection holds for *strictly positive*, or *regular* probability distributions)
- Intuitive paraphrases of these properties are as follows (Pearl)
  - Symmetry: In any state of knowledge Z, if Y tells us nothing about X, then X tells us nothing about Y.
  - Decomposition: If two combined items of information are judged irrelevant to X, then each separate item is irrelevant as well.
  - Weak Union: learning irrelevant information W cannot help the irrelevant information Y become relevant to X.
  - Contraction: If we judge W irrelevant to X after learning some irrelevant information Y, then W must have been irrelevant before we learned Y.[a]
  - Intersection: If Y is irrelevant to X when we know W and if W is irrelevant to X when we know Y, then neither W nor Y (nor $W \cup Y$) is relevant to X.
- If $(X \perp\!\!\!\perp Y \mid \emptyset)$, then X and Y are *unconditionally independent* under Pr.

[a] Weak union + contraction imply that irrelevant information should not alter the status of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant.

Markovian Parents and the Calculation of $\Pr(v)$

- Specifying a joint distribution $\Pr(v)$ over a set of dichotomous variables V is daunting. It requires knowing Pr over 2^{n-1} (out of 2^n) state descriptions over V.
- We can always represent Pr as a product of conditional distributions:

$$\Pr(v_1 \,\&\, \cdots \,\&\, v_n) = \prod_j \Pr(v_j \mid v_1 \,\&\, \cdots \,\&\, v_{j-1})$$

- Now, suppose that the conditional probability of V_j is not sensitive to all of its predecessors, but only depends on a small subset of its predecessors PA_j.
- PA_j are the *Markovian parents* (or *parents*, for short) of V_j, and:

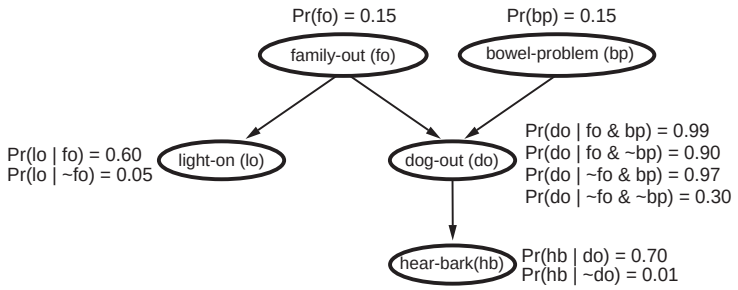
$$\Pr(v_j \mid v_1 \,\&\, \cdots \,\&\, v_{j-1}) = \Pr(v_j \mid pa_j)$$

  which *considerably simplifies* our calculation of $\Pr(v)$, above.
- More carefully, we define the *Markovian parents* PA_j of V_j in V, as the *minimal* set of predecessors of V which screen V_j off from the rest of its predecessors.

DAGs as Representers of Conditional Independence Structures

- Using the Markovian parents concept, we can construct DAGs to (uniquely) represent conditional independence structures. For any variable V_j in an ordered set V of variables, we can construct a DAG capturing PA_j, as follows:
  - Starting with the pair $\{V_1, V_2\}$, we draw an arrow from V_1 to V_2 iff the two variables are dependent [*i.e.*, if it is *not* the case that $(V_1 \perp\!\!\!\perp V_2 \mid \emptyset)$].
  - Continuing to V_3, we draw no arrow in case $(V_3 \perp\!\!\!\perp \{V_1, V_2\} \mid \emptyset)$; otherwise, we examine whether $(V_3 \perp\!\!\!\perp V_1 \mid V_2)$ or $(V_3 \perp\!\!\!\perp V_2 \mid V_1)$.
  - In the first case $(V_3 \perp\!\!\!\perp V_1 \mid V_2)$, we draw an arrow from V_2 to V_3; in the second $(V_3 \perp\!\!\!\perp V_2 \mid V_1)$, we draw an arrow from V_1 to V_3.
  - If *neither* $(V_3 \perp\!\!\!\perp V_1 \mid V_2)$ *nor* $(V_3 \perp\!\!\!\perp V_2 \mid V_1)$, then draw $V_1 \rightarrow V_3$ *and* $V_2 \rightarrow V_3$.
  - In general, at the j^{th} stage of the construction, we select a minimal set of V_j's predecessors that screens-off V_j from its other predecessors. We call this set PA_j, and draw an arrow from each $V \in PA_j$ to V_j. This generates a DAG.
- The result is a *Bayesian network*: $V_i \rightarrow V_j$ iff V_i is a Markovian parent of V_j.

An Example Bayesian Network (Charniak)

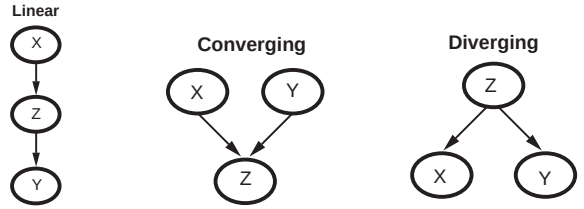


Suppose when I go home at night, I want to know if my family is home before I try the doors. (Perhaps the most convenient door to enter is double locked when nobody is home.) Now, often when my wife leaves the house, she turns on an outdoor light. However, she sometimes turns on this light if she is expecting a guest. Also, we have a 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, I will probably hear her barking (or what I think is her barking), but sometimes I can be confused by other dogs barking. This might yield the graph above.[a]

[a]Note how we only need 10 probabilities to determine $\Pr(v)$. This is *a lot* better than $2^5 = 32$!

d-connectedness and d-separation

- We want a procedure for determining whether $(V_1 \perp\!\!\!\perp V_2 \mid Z)$ in $\mathcal{G}$ $[V_1, V_2 \notin Z]$.
- There are types of connections between triples of variables X, Y, and Z.



- A path from V_1 to V_2 is *d-connected* by (a *set*) Z in $\mathcal{G}$ iff either:
  1. It is linear or diverging and does not contain any member of Z.
  2. It is converging, and some interior node on the path (or one of its descendants) is in Z.

  – If Z does not d-connect any path from V_1 to V_2 in $\mathcal{G}$, then V_1 *and* V_1 *are d-separated by* Z *in* $\mathcal{G}$, *and* $(V_1 \perp\!\!\!\perp V_2 \mid Z)$, *on any* Pr *consistent with* $\mathcal{G}$.

Applying *d*-connectedness and *d*-separation to Charniak's Example

- We have the following facts concerning Charniak's example:
  1. (family-out ⫫ hear-bark | dog-out), because (i) all paths from family-out to hear-bark contain dog-out, but (ii) none of these paths is converging.
  2. The same goes for (bowel-problem ⫫ hear-bark | dog-out)
  3. But, {dog-out} does *not d*-separate family-out from bowel-problem. To see that {dog-out} *d*-connects family-out and bowel-problem, note that disjunct (2) is satisfied, since dog-out is the only member of {dog-out}, and the path from family-out to bowel-problem is *converging* at dog-out.
  4. Similar results can be proven involving the other combinations.
- There are many software packages which will calculate the full joint distribution over a Bayesian network, from the Markovian specification of probabilities.
- These programs also allow for calculating the new joint probability distribution generated by "observing" the states of any combination of nodes in the network. Audrey Yap has written a nice JAVA Applet that does these calculations...