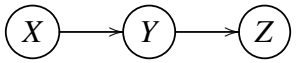
NETWORK ANALYSIS

Lourens Waldorp (adjusted by Claudia van Borkulo)

PROBABILITY AND GRAPHS

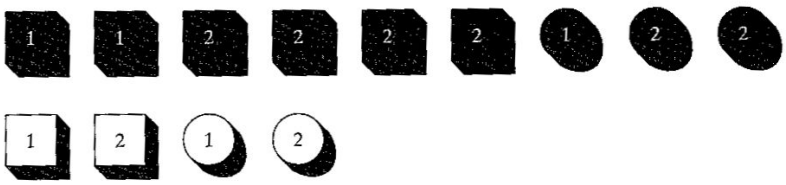
The objective is to obtain a correspondence between the intuitive pictures (graphs) of variables of interest and the probability distributions of the variables. That means that if we had a graph like this



and X, Y and Z are variables, then we would immediately see what the causal relations are in terms of the probability distributions. For that we need to know

1. what conditional independencies are implied by the graph
2. whether these independecies correspond to the probability distribution

EXAMPLE
Throughout these notes we will use a single example as a case study to apply to all definitions we introduce. Consider these objects



There are thirteen separate things although several are the same with respect to value (1 or 2), shape (square or circle), and color (black or white). Let's use Laplace's principle of indifference to set the probability of obtaining any object at 1/13.

AXIOMATIC PROBABILITY

Probability can be interpreted in several ways: frequentist, Bayesian, propensity. However, from a mathematical point of view, i.e. not considering semantics, probability can be defined in terms of a function that satisfies certain conditions, called axioms. These axioms were defined by Kolmogorov in 1933 as part of a solution to one of Hilberts famous 23 problems in mathematics from his address in 1900 in Paris (Hilbert, 1900).

Kolmogorov axioms (finite version)
Let Ω be a sample space with n distinct elements elements

$$\Omega = \{x_1, x_2, \ldots, x_n\}.$$

An event X is a subset of Ω; $X = \{x_1, x_3\}$ for example. Two events X and Y are disjunctive iff its intersection is $X \cap Y = \phi$. A function P that assigns to each event $X \in \Omega$ a real number is called a probability function on the set of subsets of Ω iff for P

(1) $0 \leq P(X) \leq 1$

(2) $P(\Omega) = 1$

(3) for X and $Y \subset \Omega$ such that $X \cap Y = \phi$

$$P(X \cup Y) = P(X) + P(Y)$$

If an event contains a single element x_i then this event is an elementary event. And so, if $X = \{x_1\}$ and $Y = \{x_3\}$ are elementary events, then $X \cap Y = \phi$ and

$$P(X \cup Y) = P(X) + P(Y) = P(\{x_1\}) + P(\{x_3\})$$

Example. These axioms can be verified for our sample space of objects. In our example an elementary event is a single object. So if we number each element in Fig. from x_1 to x_{13}, then $\Omega = \{x_1, \ldots, x_{13}\}$. Since two elementary events are disjunctive, that is the objects are different things, so $\{x_1\} \cap \{x_2\} = \phi$. Then we have by (3)

$$P(x_1 \cup x_2) = P(x_1) + P(x_2) = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}.$$

And since

$$\Omega = \bigcup_{i=1}^{13} x_i$$

we have that

$$P(\Omega) = P(\cup_{i=1}^{13} x_i) = 1.$$

Let's define three events for our example: X for value, Y for shape, and Z for color. That is

$$\begin{aligned} X = 1 &: \text{ all objects containing 1} \\ X = 2 &: \text{ all objects containing 2} \\ Y = s &: \text{ all square objects} \\ Y = c &: \text{ all circular objects} \\ Z = w &: \text{ all white objects} \\ Z = b &: \text{ all black objects} \end{aligned}$$

If we consider $X = 1$, then we have 5 objects (disregarding shape and color) out of thirteen. So

$$P(X = 1) = \frac{5}{13}.$$

A probability of a combination of events, a conjunction, is called a joint probability. For example, the probability of $\{X = 1, Y = s, Z = b\}$ is a joint probability and is in this case

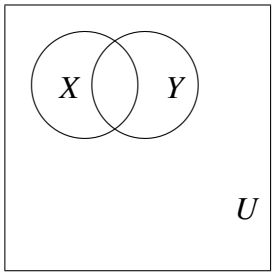
$$P(\{X = 1, Y = s, Z = b\}) = \frac{2}{13}.$$

It is then immediately clear what we did to get to the probability of $\{X = 1\}$, we disregarded, or collapsed, or marginalized over the other two variables X and Z. So,

$$\begin{aligned} P(X = 1) &= \sum_{Y=s,c} \sum_{Z=w,b} P(X = 1, Y = y, Z = z) \\ &= P(X = 1, Y = s, Z = w) + P(X = 1, Y = c, Z = w) \\ &\quad + P(X = 1, Y = s, Z = b) + P(X = 1, Y = c, Z = b) \\ &= \frac{1}{13} + \frac{1}{13} + \frac{2}{13} + \frac{1}{13} \\ &= \frac{5}{13}. \end{aligned}$$

Conditional probability and independence

Conditional probability satisfies itself Kolmogorov's axioms and so is itself a probability function. It can therefore be considered as a probability with a 'new' reference universe. The new universe is the limitation of the variable of interest to another variable, the variable that is being conditioned on. This can be seen in the Venn diagram below.



If Y were conditioned on, then Y is the new universe.

Conditional probability
For two events X and Y, the conditional probability of $X = x$ given $Y = y$, for $P(y) > 0$, is defined as

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

Example. Consider the same events X and Y as above. If we are interested in the event $X = 1$ given that $Y = s$, then we require the joint probability of $X = 1$ and $Y = s$ and the probability of $Y = s$. The probability of $Y = s$ is the number of square objects divided by the number of objects, that is

$$P(Y = s) = \frac{8}{13}.$$

The conjunction of $X = 1$ and $Y = s$ refers to the objects that are both valued 1 and are square. So

$$P(X = 1, Y = s) = \frac{3}{13}.$$

Then

$$P(X = 1 \mid Y = s) = \frac{3/13}{8/13} = \frac{3}{8}.$$

It can be seen here that the conditional probability considers the possible objects of $Y = s$ as the new universe. $Y = s$ contains 8 elements, all objects that are square disregarding the other two features.

The conditional probability behaves like a normal probability (and in fact is a probability function). One of the properties is that marginalizing (summing over all possible values of) the variable of interest, will yield a probability of 1, as in a unconditional probability. Suppose we marginalize over X given that $Y = s$. Then

$$\begin{aligned}\sum_{X=1,2} P(X \mid Y = s) &= \frac{P(X = 1, Y = s)}{P(Y = s)} + \frac{P(X = 2, Y = s)}{P(Y = s)} \\ &= \frac{3/13}{8/13} + \frac{5/13}{8/13} = \frac{3}{8} + \frac{5}{8} = 1\end{aligned}$$

Chain rule
The joint probability of 3 variables X, Y, and Z, $P(x, y, z)$, permits the factorization

$$P(x, y, z) = P(x)P(y \mid x)P(z \mid x, y)$$

This is easily seen to be true, because, using the definition of conditional distribution

$$P(x)P(y \mid x)P(z \mid x, y) = P(x)\frac{P(y, x)}{P(x)}\frac{P(z, x, y)}{P(x, y)} = P(x, y, z)$$

The chain rule can be shown to hold for any number n of random variables by induction on n.

Example. Suppose we would like to know the joint probability of $X = 1$, $Y = c$, and $Z = w$, then we know by counting the objects that

$$P(X = 1, Y = c, Z = w) = \frac{1}{13},$$

because there is one such object. Using the (conditional) probabilities

$$\begin{aligned} P(X = 1) &= \frac{5}{13} \\ P(Y = c \mid X = 1) &= \frac{2/13}{5/13} \\ P(Z = w \mid X = 1, Y = c) &= \frac{1/13}{2/13} \end{aligned}$$

we have

$$P(X = 1)P(Y = c \mid X = 1)P(Z = w \mid X = 1, Y = c) = \frac{5}{13}\frac{2}{5}\frac{1}{2} = \frac{1}{13}$$

Bayes rule
For two events X and Y with neither $P(x) = 0$ nor $P(y) = 0$, the inverse probability of $P(y \mid x)$ is

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}, \quad P(y) = \sum_{X=x} P(y, x)$$

Example. Suppose we would like to know the joint probability of conditional probability of $X = 1$ given $Y = c$, but we have the inverse of this. That is, we know that $P(Y = c \mid$

$X = 1) = 2/5$ and $P(X = 1) = 5/13$, then we can use Bayes rule to get $P(X = 1 \mid Y = c)$. This is done by

$$P(X = 1 \mid Y = c) = \frac{P(Y = c \mid X = 1)P(X = 1)}{P(Y = c)} = \frac{(2/5)(5/13)}{5/13} = \frac{2}{5}$$

Note that $P(Y = c) = \sum_{x=1,2} P(Y = c, X = x)$ and that $P(Y = c, X = 1) = P(Y = c \mid X = 1)P(X = 1)$, so together with $P(Y = c \mid X = 2)$ and $P(X = 2)$ all information is available.

Independence
Two events X and Y are (statistically) independent iff one of the following holds

1. if $P(x) \neq 0$ and $P(y) \neq 0$ then $P(x \mid y) = P(x)$.
2. $P(x) = 0$ or $P(y) = 0$.

The first part is the interesting one, the second part is required for completeness (why?). In words, independence means that the probability of $X = x$ occurring does not depend on the value of Y. Given the definition of conditional probability, it follows directly that if $X = x$ and $Y = y$ are independent, then

$$P(x) = P(x \mid y) = \frac{P(x, y)}{P(y)} = \frac{P(x)P(y)}{P(y)}$$

Example. The events $X = 1$ and $Y = s$ are not independent because

$$P(X = 1 \mid Y = s) = \frac{3/13}{8/13} = \frac{3}{8}$$
$$P(X = 1) = \frac{5}{13}$$

The concept of independence can be extended to include more than two variables. This is referred to as conditional independence.

Conditional independence
The events X and Y are conditionally independent given Z with $P(z) \neq 0$ iff one of the following holds

1. $P(x \mid y, z) = P(x \mid z)$ and $P(x \mid z) \neq 0$ and $P(y \mid z) \neq 0$
2. $P(x \mid z) = 0$ or $P(y \mid z) = 0$

Example. The events $X = 1$ and $Y = s$ are conditionally independent given $Z = b$. This can be verified by considering the different probabilities according to the definition

$$P(X = 1 \mid Z = b) = \frac{3}{9} = \frac{1}{3}$$
$$P(X = 1 \mid Y = s, Z = b) = \frac{2}{6} = \frac{1}{3}$$

The principle of a new universe still applies, as you can see from the example. Only now the new universe in the conditional probability is determined by two (or more in general) variables.

To determine whether the variables X and Y are conditionally independent given Z we need to determine the probabilities for all values each of X, Y and Z can assume. So, we have a table of probabilities as follows

x	y	z	$P(x \mid y, z)$	$P(x \mid z)$
1	s	b	1/3	1/3
2	s	b	2/3	2/3
1	c	b	1/3	1/3
2	c	b	2/3	2/3
1	s	w	1/2	1/2
2	s	w	1/2	1/2
1	c	w	1/2	1/2
2	c	w	1/2	1/2

So, knowledge of being square or circle is irrelevant to whether the object has value 1 or 2 given that you know that its color is black or white. Conditional independence has become very important in statistics and has its own notation, due to Dawid. If two events X and Y are conditionally independent given Z then this is written as

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{iff} \quad P(x \mid y, z) = P(x \mid z)$$

The subscript P is there to indicate that the conditional independence is true in the probability distribution. To show conditional independence it is sometimes convenient to rewrite the probability statement.

$$P(x \mid y, z) = P(x \mid z) \quad \Leftrightarrow \quad P(x, y \mid z) = P(x \mid z)P(y \mid z)$$

GRAPHS

A graph is a combination of two sets: a set V of vertices or nodes and a set E of edges

or links. The set V contains variables; in the case of probabilistic inference it contains random variables. These could, for instance, be the set of random variables from the example, X, Y, and Z. The set E contains ordered pairs indicating a link between two variables in V. For instance, (X, Y) refers to an arrow (directed edge) from X to Y, as in $X \rightarrow Y$. Two examples of directed graphs are given in Fig 1. A directed path is a set of edges that follows the direction of the edges. In the graph G_2 of Fig 1, for example, a directed path is the set $\{(X, Y), (Y, Z)\}$.

Directed acyclic graph
The combination (V, E) is referred to as a directed acyclic graph (DAG) iff

(i) there are only directed edges (arrows) in E, and

(ii) there is no directed path such that the first node is the same as the last, i.e. there are no loops.

For example, the graph G_2 with $V = \{X, Y, Z\}$ is cyclic if its edge set would also contain (Z, X).

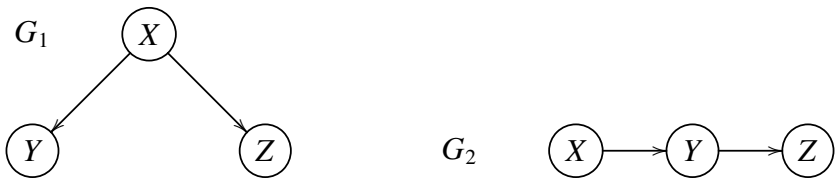


Figure 1: Two DAGs G_1 and G_2 with edge sets $E_1 = \{(X, Y), (X, Z)\}$ and $E_2 = \{(X, Y), (Y, Z)\}$. Behalve hier en in de paragraaf hieronder wordt er daarna gerefereerd naar het voorbeeld X $\rightarrow$ Z $\rightarrow$ Y. Misschien beter als ook hier en in paragraaf hieronder uitgaan van X $\rightarrow$ Z $\rightarrow$ Y?

Kinship is used to refer to topological structure in graphs. So, the parents refer to adjacent nodes at the tail end of the arrows. In G_1 the parent of Y is X, and in G_2 the parent of Z is Y. Descendants are nodes that are at the head end of the arrow. So, in G_1, Y is a descendant (and a child) of X and Z is also a descendant of X (but not a child).

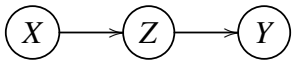
GRAPHICAL MODELS: COMBINING PROBABILITY DISTRIBUTIONS AND GRAPHS
In the example, we established already that X and Y are independent given Z, that is we have that

$$(X \perp\!\!\!\perp Y \mid Z)_P.$$

From the definition of the notation, we know that the joint distribution of the three variables is

$$P(x, y, z) = P(x)P(z \mid x)P(y \mid z) \quad \Leftrightarrow \quad P(y \mid x, z) = P(y \mid z)$$

What we want is a graph to represent this conditional independence relation between variables. If there is correspondence, then this conditional independence can be read off from a graph, like this one



In that case we write

$$(X \perp\!\!\!\perp Y \mid Z)_G,$$

where the subscript G now indicates that the graph represents the conditional independence. So, now we associate the independence in the probability distribution $(X \perp\!\!\!\perp Y \mid Z)_P$ with $(X \perp\!\!\!\perp Y \mid Z)_G$. The probability statement means that the table that was computed for each of the $2^3 = 8$ combinations of possible values for X, Y, and Z shows this independence relation to be true. The graph says the same intuitively. But we are not yet there. We need certain conditions for the exact correspondence between a graph and a probability distribution to hold. What we need is

1. an arrow implies dependence
2. absence of an arrow implies conditional independence when knowing only the parents
3. no other conditional independencies can exist other than those implied by the arrows

Condition 2 refers to what is called the Markov condition. We know from the chain rule that any joint probability distribution can be written in terms of conditional distributions, that is

$$P(x, y, z) = P(x)P(z \mid x)P(y \mid x, z)$$

The Markov condition says that we can reduce the conditional part to only the parents. So, in this case we would get

$$P(x, y, z) = P(x)P(z \mid x)P(y \mid z)$$

This may not look like much, but that is because it is a small network. This is of course the conditional independence $(X \perp\!\!\!\perp Y \mid Z)_P$.

Markov condition
A probability distribution is Markov relative to a DAG $G = (V, E)$ with $V = \{X, Y, Z\}$ iff

for each $U \in V$, U is independent of its nondescendants nd_u (without U itself) given the parents pa_u of U. So,

$$(X \perp\!\!\!\perp nd_x \mid pa_x)_P \Leftrightarrow P(x \mid y, z) = P(x)$$
$$(Z \perp\!\!\!\perp nd_z \mid pa_z)_P \Leftrightarrow P(z \mid x) = P(z \mid x)$$
$$(Y \perp\!\!\!\perp nd_y \mid pa_y)_P \Leftrightarrow P(y \mid x, z) = P(y \mid z)$$

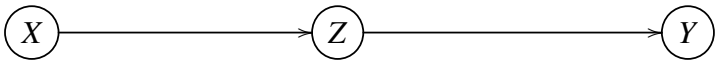
The important consequence is that if the pair G and P is Markov, then we have the easy factorization.

Markov compatibility
If P is Markov relative to G, then the joint probability can be written as the product of conditional distributions of all variables given the parents. And so

$$P(x, y, z) = P(x)P(z \mid x)P(y \mid z)$$

We now have a correspondence between a DAG and a probability distribution. And we can get from the conditional probabilities to the joint distribution.

Example. Consider the graph G_1 that represents P for the 13 objects and their corresponding conditional probabilities



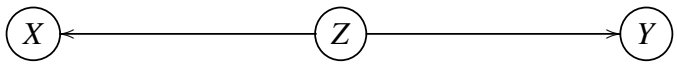
$$P(X = 1) = \tfrac{5}{13} \qquad P(Z = b \mid X = 1) = \tfrac{3}{5} \qquad P(Y = s \mid Z = b) = \tfrac{2}{3}$$

$$P(Z = b \mid X = 2) = \tfrac{3}{4} \qquad P(Y = s \mid Z = w) = \tfrac{1}{2}$$

Then the joint probability is the product of the conditional probabilities if P is Markov with respect to G_1. So,

$$\begin{aligned} P(X = 1, Y = s, Z = b) &= \frac{2}{13} \\ &= P(X = 1)P(Z = b \mid X = 1)P(Y = s \mid Z = b) \\ &= \frac{5}{13}\frac{3}{5}\frac{2}{3} \end{aligned}$$

If you finish this list for all 2^3 possible combinations of X, Y, and Z, then you can conclude that the DAG G_1 is a representation of P and so that it is Markov relative to G_1. However, there is nothing to say that this is the only DAG that leads to a product of conditional distributions which constitute the joint probability P. The DAG G_2 has the following properties



$$P(X = 1 \mid Z = b) = \tfrac{1}{3} \qquad P(Z = b) = \tfrac{9}{13} \qquad P(Y = s \mid Z = b) = \tfrac{2}{3}$$

$$P(X = 1 \mid Z = w) = \tfrac{1}{2} \qquad P(Y = s \mid Z = w) = \tfrac{1}{2}$$

For these conditional distributions, you can see that it is also Markov compatible, because

$$\begin{aligned} P(X = 1, Y = s, Z = b) &= \frac{2}{13} \\ &= P(Z = b)P(X = 1 \mid Z = b)P(Y = s \mid Z = b) \\ &= \frac{9}{13}\frac{1}{3}\frac{2}{3} \end{aligned}$$

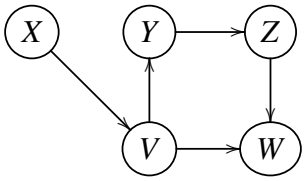
So there is no way to decide between these two DAGs based on the information available to us now. These DAGs are both equivalent representations of the joint probability distribution P of the objects.

An important way to get to the conditional independencies is by considering how 'information' from nodes is transferred across the DAG. If it is possible to determine a way to read off the DAG how information can be blocked by certain nodes, then it may be possible to determine which variables are conditionally independent in the distribution. This is done by a criterion called d-separation (d for dependent). This provides a strong procedure to determine conditional independencies by reading DAGs.

d-separation
Two disjoint sets of variables X and Z are d-separated by another disjoint set Y iff either

1. a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ with m in the set Y; or
2. there is a collider $i \to m \leftarrow j$ with m nor any descendants in Y

Different example. Suppose we have the following DAG



To determine the d-separations in the DAG we need to see whether the path is blocked for each path between variables.

d-separation	path	value
$(X \perp\!\!\!\perp Z \mid Y)_G$	$X \to V \to W \leftarrow Z$	blocked
	$X \to V \to Y \to Z$	blocked
$(X \perp\!\!\!\perp Z \mid YW)_G$	$X \to V \to Y \to Z$	blocked
	$X \to V \to W \leftarrow Z$	active

So, the d-separation $(X \perp\!\!\!\perp Z \mid Y)_G$ implies the corresponding conditional independence $(X \perp\!\!\!\perp Z \mid Y)_P$. But $(X \perp\!\!\!\perp Z \mid YW)_G$ does not imply $(X \perp\!\!\!\perp Z \mid YW)_P$ because one of the paths is active.

Theorem 1.2.4 and 1.2.5
These ideas are generalized in theorems 1.2.4 and 1.2.5 in Pearl (2000, p. 18). It says that if all d-separations are found then all distributions that are Markov compatible will have these conditional independencies. And if all distributions have these conditional independencies then the DAG will have these d-separations. So we can go back and forth between DAG and probability distribution.

Remark. These two theorems do not imply that all causal relations can be derived by considering data alone. Indeed, on pages 10 – 11 we saw that two causally different graphs were both Markov compatible to the same probability distribution. In general, the three models in the top row are equivalent and the one in the second row can be distinguished from the other three because it is a collider.