BCCS 2008/09: Graphical models and complex stochastic systems: Lectures 3/4: Conditional independence and graphs

After that general scene-setting and introduction, today we begin to turn to what is special about this module – the idea of **structure** in statistical model-building.

3.1 Conditional independence

The kind of structure we are talking about is specified in terms of **conditional independence**, and let's get to that by first talking about ordinary independence. Suppose we have two discrete random variables X and Y; they are said to be independent if for all X and Y, $P\{X = y, Y = y\} = P\{X = x\} \times P\{Y = y\}$, or using the shorthand notation from Lecture 2, $p(x, y) = p(x)p(y)$. (In this form the definition applies to continuous random variables too). Two things should be stressed: (1) this is a model assumption that may or may not be scientifically reasonable in a particular situation; (2) it is a property of the probability not the random variables – two people may assign probabilities differently to the random variables (X, Y), and one may make them independent, the other not. For example, let (X, Y) be the coordinates of the random point where a dart lands on a board. For a thrower whose aiming error pattern is 'circular', X and Y will be (modelled as) independent, but for one with a NE/SW bias, they will be dependent.

Independence is symmetric between X and Y. To understand the term, consider the conditional probability $p(y|x)$ (short for $P\{Y = y|X = x\}$). By definition, $p(y|x) = p(x, y)/p(x)$, so X and Y are independent if and only if $p(y|x) = p(y)$, read as saying that 'knowing x tells you nothing about y'. Hence, 'independence'. Note that if x tells you nothing about y, then y tells you nothing about x.

You can see that conditional probability and independence begins to provide a language for describing whether some variables (e.g. data) provide any information about other variables (e.g. parameters).

To talk about conditional independence, we need at least three random variables, say X, Y and Z. We say X and Y are **conditionally independent** given Z if $p(x, y|z) = p(x|z) \times p(y|z)$, which is the same as $p(x|z, y) = p(x|z)$: in words, 'if you know z, then learning y tells you nothing new about x'. We write this for short as $x \perp\!\!\!\perp y \mid z$.

Conditional independence satisfies various rules, which all make sense intuitively, and can also be checked mathematically very simply:

1. if $x \perp\!\!\!\perp y \mid z$ then $y \perp\!\!\!\perp x \mid z$
2. if $x \perp\!\!\!\perp y \mid z$ and $u = h(x)$ then $u \perp\!\!\!\perp y \mid z$
3. if $x \perp\!\!\!\perp y \mid z$ and $u = h(x)$ then $x \perp\!\!\!\perp y \mid (z, u)$
4. if $x \perp\!\!\!\perp y \mid z$ and $x \perp\!\!\!\perp w \mid (y, z)$ then $x \perp\!\!\!\perp (w, y) \mid z$
5. *Under extra conditions, e.g.* $p(x, y, z, w) > 0 \forall (x, y, z, w)$, if $x \perp\!\!\!\perp y \mid (z, w)$ and $x \perp\!\!\!\perp z \mid (y, w)$ then $x \perp\!\!\!\perp (y, z) \mid w$.

Perhaps surprisingly, these rules can actually be used as a set of axioms to define conditional independence as a free-standing concept, without referring to probability at all. We won't explore that idea.

Example

Suppose we give a particular drug treatment to n patients with a certain medical condition. Let the random variable X be the number who respond positively, and let Y be 1 or 0 according to whether the first patient *tomorrow* will respond positively. Are X and Y independent? (Does X tell you anything about Y?) Compare with coin-tossing: is that similar or different? Now let θ be the probability that a randomly chosen patient will respond positively. What we probably want to assume is *not* that X and Y are independent, but that they are *conditionally* independent given θ: $X \perp\!\!\!\perp Y \mid \theta$.

Graphical modelling

Graphical modelling is concerned with representing the conditional independence relations among a collection of random variables graphically. A graph is a diagram consisting of points (called nodes or vertices), joined by lines (called arcs or edges), which may or may not have arrowheads indicating direction. In graphical modelling, the vertices represent random variables, and the edges collectively indicate the conditional independence properties.

As we shall see, these graphical representations assist in visualisation and interpretation, for input and output from computer systems, in constructing algorithms, and in mathematical proofs.

3.2 Directed acyclic graphs

There are several kinds of graph used in this subject, but we will first concentrate on *directed acyclic graphs* (DAGs), that is graphs in which all edges are directed, and there are no (directed) loops (in particular, no edges from a vertex to itself).

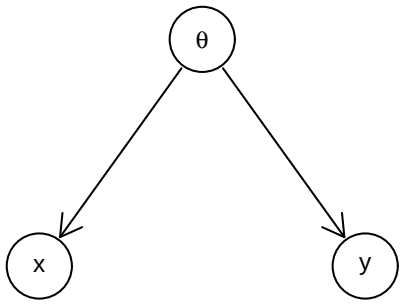
Parents and children. One important use for DAGs is to models of the genes of related individuals. A standard assumption is that, given its parents' genotypes, an individual's genotype is independent of those of the grandparents (and of brothers, sisters, aunts, ...)

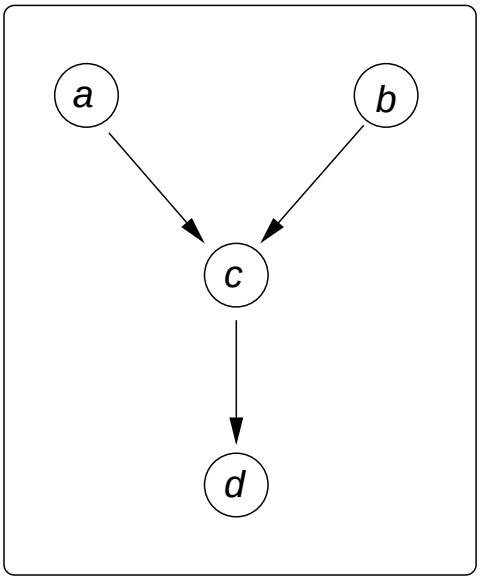
We borrow the terminology more generally: the variables from which an arrow leads to variable x are the *parents* of X, denoted $\mathrm{pa}(x)$. Those x for which $y \in \mathrm{pa}(x)$ are the *children* of y. Variables with no parents are called *founders*.

With a vector x of variables indexed by v, we will often write $\mathrm{pa}(v)$ rather than $\mathrm{pa}(x_v)$ for simplicity. We also write x_A for the sub-vector consisting of the components indexed by $v \in A$, so $x_{\mathrm{pa}(v)}$ stands for $\{x_w : w \in \mathrm{pa}(x_v)\}$.

Coin-tossing example.

This graph represents $x \perp\!\!\!\perp y \mid \theta$, the conditional independence of x and y given θ. We have $p(x, y|\theta) = p(x|\theta)p(y|\theta)$, so $p(\theta, x, y) = p(\theta)p(x|\theta)p(y|\theta)$.





This graph is drawn to symbolise the statement that $p(a,b,c,d) = p(a)p(b)p(c|a,b)p(d|c)$. Since it is always true that $p(a,b,c,d) = p(a)p(b|a)p(c|a,b) \times p(d|a,b,c)$, it indicates that we are assuming $a \perp\!\!\!\perp b$ and $d \perp\!\!\!\perp (a,b) \mid c$.

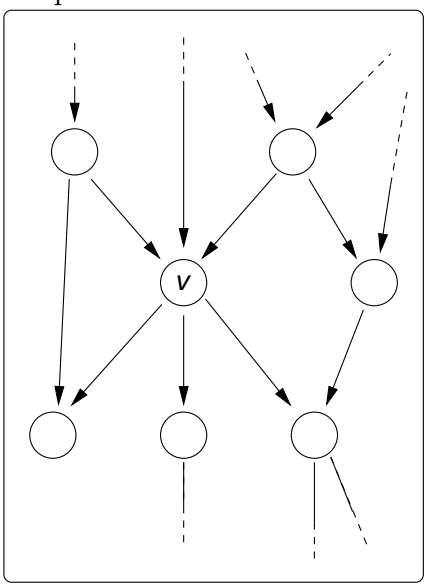
It is always true that

$$p(x_1, x_2, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_1, x_2, \ldots, x_{i-1})$$

but it is useful and relevant to build models in which not all of $\{x_1, x_2, \ldots, x_{i-1}\}$ are involved in the condition in the ith factor.

The DAG corresponds to a way of factorising $p(x_1, x_2, \ldots, x_n)$: we draw arrows to indicate which variables are needed in each factor: e.g. if the 'd' factor is $p(d|c)$ not $p(d|a,b,c)$ it means we do not have arrows from a or b to c.

Thus absence of arrows indicates conditional independence.



A more general directed acyclic graph (DAG). It symbolises the fact that the joint distribution of all variables factorises: $p(x) = \prod_v p(x_v | x_{\mathrm{pa}(v)})$.

3.3 Markov properties for DAGs

Earlier we commented in a genetics context that, given its parents' genotypes, an individual genotype was independent of certain others.

This is an example of a Markov property, analogous to the property in a Markov chain that $x_{n+1} \perp\!\!\!\perp (x_{n-1}, x_{n-2}, \ldots) \mid x_n$.

The **local Markov property** for a DAG is that each individual variable is independent of its *non-descendants*, given its parents, i.e.

$$x_v \perp\!\!\!\perp x_{\mathrm{nd}(v)} \mid x_{\mathrm{pa}(v)}.$$

The non-descendants are all the variables, except for x_v, its parents, and its descendants (those w for which v is a parent of a parent of ... of w).

Proof. Partition the variables into 4 disjoint subsets: $\{v\}$, $\mathrm{pa}(v)$, $\mathrm{nd}(v)$ and the descendants of v, denoted $\mathrm{de}(v)$. We know $p(x) = \prod_w p(x_w|x_{\mathrm{pa}(w)})$: write this as the product of 4 products:

$$p(x_v|x_{\mathrm{pa}(v)}) \prod_{w\in\mathrm{pa}(v)} p(x_w|x_{\mathrm{pa}(w)}) \prod_{w\in\mathrm{nd}(v)} p(x_w|x_{\mathrm{pa}(w)}) \prod_{w\in\mathrm{de}(v)} p(x_w|x_{\mathrm{pa}(w)})$$

Marginalise (sum or integrate) out all the variables in $\mathrm{de}(v)$ – the 4th term becomes 1. Now note that x_v or its descendants cannot appear in any of the factors in the 2nd term (by the acyclicity of a DAG) or 3rd term (by definition of non-descendant). So the 2nd and 3rd terms are functions only of $x_{\mathrm{pa}(v)}$ and $x_{\mathrm{nd}(v)}$. Thus $p(x_v, x_{\mathrm{pa}(v)}, x_{\mathrm{nd}(v)})$ is a product of functions of $(x_v, x_{\mathrm{pa}(v)})$ and of $(x_{\mathrm{pa}(v)}, x_{\mathrm{nd}(v)})$ (no term involving both x_v and $x_{\mathrm{nd}(v)}$). This means $x_v \perp\!\!\!\perp x_{\mathrm{nd}(v)} \mid x_{\mathrm{pa}(v)}$, as required.

The **global Markov property** for a DAG makes a statement about 3 sets of nodes A, B, C, say. Start with the original graph, then

1. delete all nodes that are not in A, B or C, or *ancestors* of nodes in A, B or C, and all arrows into or out of deleted nodes.

2. add an edge between each pair of parents that are not already connected (*moralisation*).

3. drop directions on all the arrows

Now look at the resulting graph to see if C *separates* A and B – that is, if it is impossible to find a path from a node in A to a node in B that does not pass through C. The global Markov property says that if C separates A and B, then $A \perp\!\!\!\perp B \mid C$.

There is an important **theorem** that says that the local and global Markov properties for a DAG are equivalent to the statement that the joint distribution factorises as required: $p(x) = \prod_v p(x_v|x_{\mathrm{pa}(v)})$.

3.4 Reading

Lauritzen, chapter 3, covers everything in a very rigorous way. Whittaker's book is a more accessible source for the role of conditional independence in data analysis. The main practical use for this material is in modelling, so that the next lectures will be useful in reinforcing these ideas.