Graphical Model Notation

Joint distributions

The joint distribution of N random variables is a very general way to encode knowledge about a system.

They can always be decomposed into a set of simpler conditional distributions by the chain rule of probability

$$p(x_1, x_2, \ldots, x_N) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)\ldots p(x_N|x_{N-1}, \cdots, x_1)$$

this is true for any joint distribution over any random variables. For example, an application of the chain rule for two random variables gives

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

and for N random variables

$$p(x_1, x_2, \ldots, x_N) = \prod_{j=1}^{N} p(x_j|x_1, x_2, \ldots, x_{j-1})$$

for all possible orderings.

This decomposition doesn't reduce the number of parameters.

Conditional Independence

Two random variables A, B are conditionally independent given a third variable C, denoted

$$X_A \perp X_B | X_C$$

if

$$\Leftrightarrow p(X_A, X_B|X_C) = p(X_A|X_C)p(X_B|X_C)$$
$$\Leftrightarrow p(X_A|X_B, X_C) = p(X_A|X_C)$$
$$\Leftrightarrow p(X_B|X_A, X_C) = p(X_B|X_C)$$

for all X_c.

Only a subset of all joint distributions respect any given conditional independence statement.

Graphical models

Probabilistic graphical models are a concise way to specify and reason about conditional independencies, without worrying about the detailed form of the distribution. There are three flavours:

- Undirected graphs
- Factor graphs
- Directed graphs

This couse will focus on directed models, since they are the most commonly encountered, and are relatively more interpretable.

Directed acyclic graphical models

A directed graphical model implies a restricted factorization of the joint distribution. In a DAG, variables are represented by nodes, and edges represent dependence.

As above, for any joint distribution of random variables $x_1, x_2, \dots, x_N$, we can write:

$$p(x_1, x_2, \dots, x_N) = \prod_{j=1}^{N} p(x_j | x_1, x_2, \dots, x_{j-1})$$

for any ordering of the nodes, where $p(x_1|x_0) = p(x_1)$.

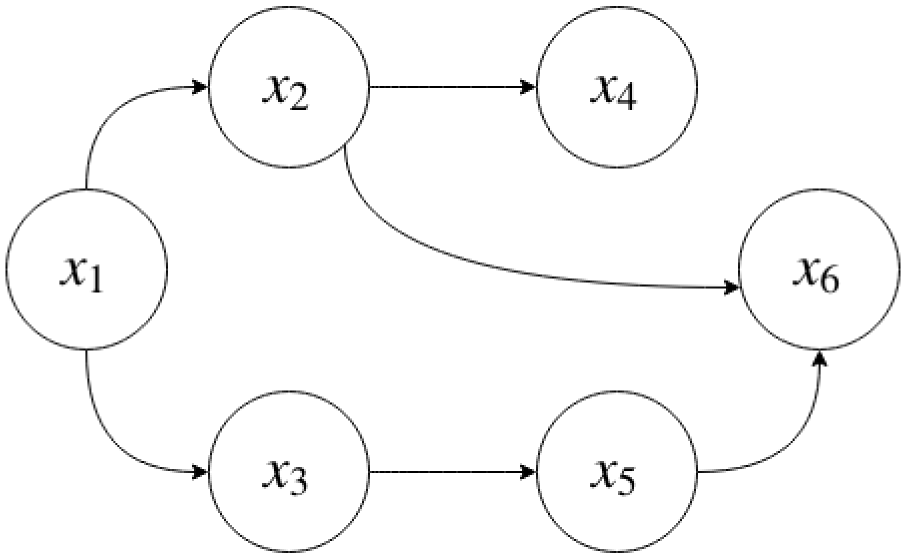
The meaning of any particular directed acyclic graphical model D is that

$$p(x_1, x_2, \dots, x_N) = \prod_{i=1}^{N} p(x_i | \text{parents}_M(x_i))$$

where $\text{parents}_M(x_i)$ is the set of nodes with edges pointing to x_i.

In other words, the joint distribution of a DAGM factors into a product of local conditional distributions, where each node (a random variable) is conditionally dependent on its parent node(s), which could be empty.

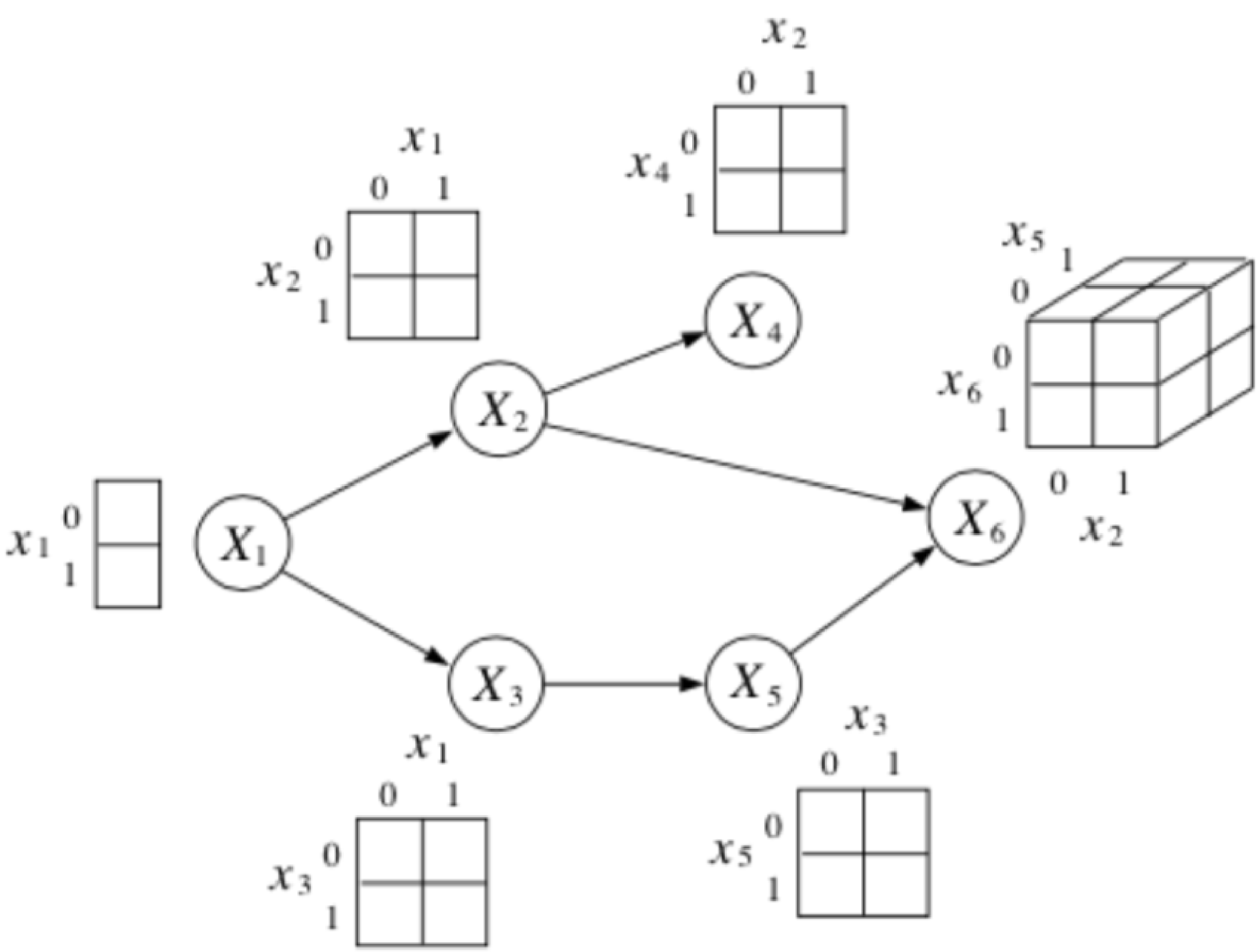
For example, the graphical model



corresponds to the following factorization of the joint distribution:

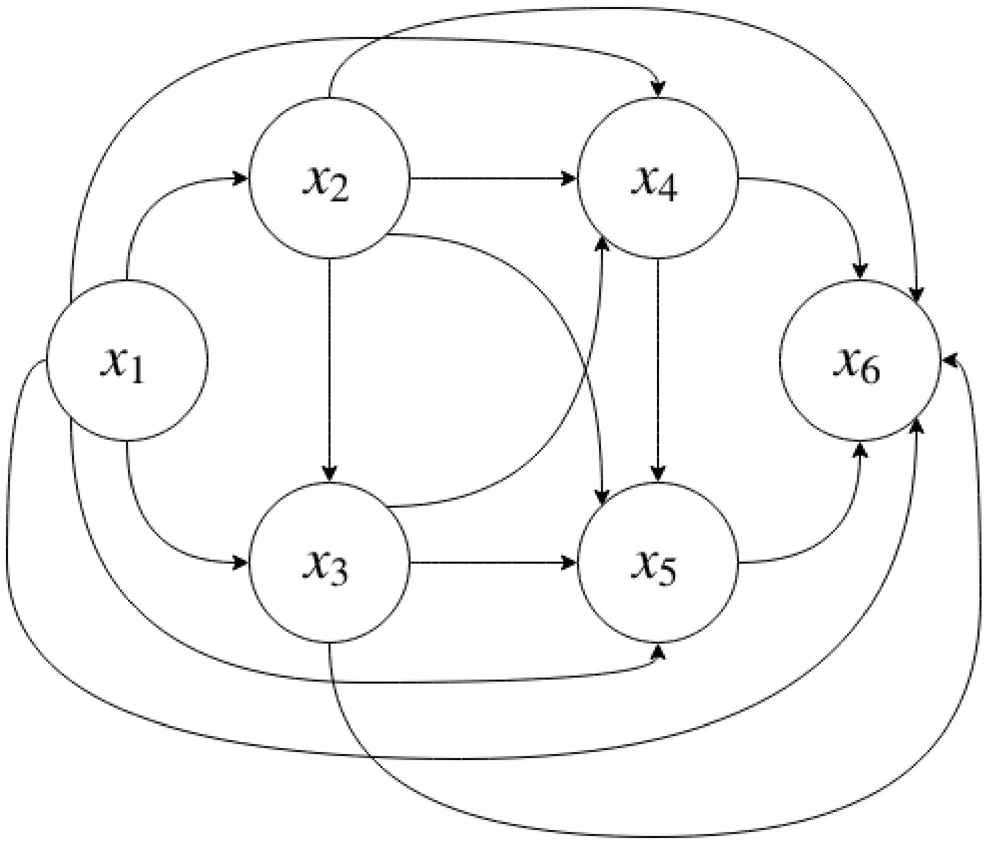
$$p(x_{1,\ldots,6}) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2,x_5)$$

Suppose each is x_i is a binary random variable.



where each conditional probabilty table with K parents requires 2^K parameters.

If we allow all possible conditional dependencies, that corresponds to a fully-connected DAG:



In general, it's computationally infeasible to work with fully-flexible distributions like this one, both because of the computational burden, and because it's hard to fit so many parameters accurately.

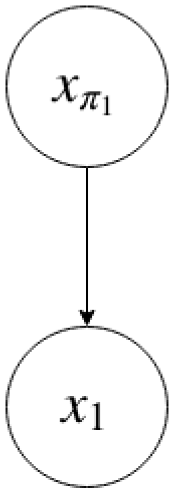
We can reduce the number of parameters in a model, and also reduce the computational burden of making inferences by introducing conditional independencies. This might not be a bad approximation in some settings.

Grouping variables

We can always group variables together into one bigger variable.

$$p(x_i, x_{\pi_i}) = p(x_{\pi_i})p(x_i|x_{\pi_i})$$

as



Conditional Independence in DAGMs

From Kevin Murphy: The simplest conditional independence relationship encoded in a Bayesian network can be stated as follows: a node is independent of its ancestors given its parents:

$$x_i \perp x_{\tilde{\pi}_i} \mid x_{\pi_i}$$

In general, missing edges *imply conditional independence*. In future lectures, we'll find out how to determine which conditional independencies hold given a DAG.