Chapter 21

Graphical Models

We have spent a lot of time looking at ways of figuring out how one variable (or set of variables) depends on another variable (or set of variables) — this is the core idea in regression and in conditional density estimation. We have also looked at how to estimate the joint distribution of variables, both with kernel density estimation and with models like factor and mixture models. The later two show an example of how to get the joint distribution by combining a conditional distribution (observables given factors; mixture components) with a marginal distribution (Gaussian distribution of factors; the component weights). When dealing with complex sets of dependent variables, it would be nice to have a general way of composing conditional distributions together to get joint distributions, and especially nice if this gave us a way of reasoning about what we could ignore, of seeing which variables are irrelevant to which other variables. This is what **graphical models** let us do.

21.1 Conditional Independence and Factor Models

The easiest way into this may be to start with the diagrams we drew for factor analysis. There, we had observables and we had factors, and each observable depended on, or loaded on, some of the factors. We drew a diagram where we had nodes, standing for the variables, and arrows running from the factors to the observables which depended on them. In the factor model, all the observables were conditionally independent of each other, given all the factors:

$$p(X_1, X_2, \ldots X_p | F_1, F_2, \ldots F_q) = \prod_{i=1}^{p} p(X_i | F_1, \ldots F_q) \tag{21.1}$$

But in fact observables are also independent of the factors they do not load on, so this is still too complicated. Let's write loads(i) for the set of factors on which the observable X_i loads. Then

$$p(X_1, X_2, \ldots X_p | F_1, F_2, \ldots F_q) = \prod_{i=1}^{p} p(X_i | F_{\text{loads}(i)}) \tag{21.2}$$

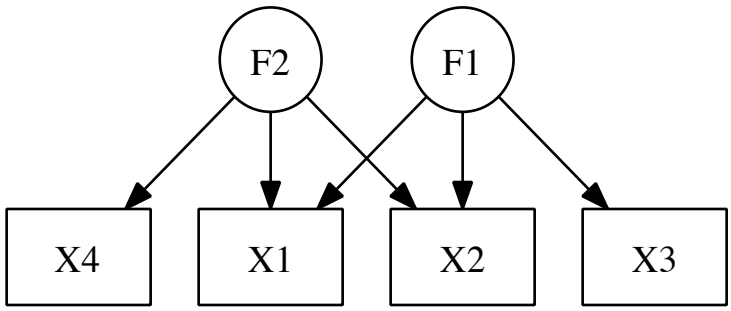

Figure 21.1: Illustration of a typical model with two latent factors (F_1 and F_2, in circles) and four observables (X_1 through X_4).

Consider Figure 21.1. The conditional distribution of observables given factors is

$$p(X_1, X_2, X_3, X_4 | F_1, F_2) = p(X_1|F_1, F_2)p(X_2|F_1, F_2)p(X_3|F_1)p(X_4|F2) \quad (21.3)$$

X_1 loads on F_1 and F_2, so it is independent of everything else, given those two variables. X_1 is unconditionally dependent on X_2, because they load on common factors, F_1 and F_2; and X_1 and X_3 are also dependent, because they both load on F_1. In fact, X_1 and X_2 are still dependent given F_1, because X_2 still gives information about F_2. But X_1 and X_3 are independent given F_1, because they have no other factors in common. Finally, X_3 and X_4 are unconditionally independent because they have no factors in common. But they become dependent given X_1, which provides information about both the common factors.

None of these assertions rely on the detailed assumptions of the factor model, like Gaussian distributions for the factors, or linear dependence between factors and observables. What they rely on is that X_i is independent of *everything else*, given the factors it loads on. The idea of graphical models is to generalize this, by focusing on relations of direct dependence, and the conditional independence relations implied by them.

21.2 Directed Acyclic Graph (DAG) Models

We have a collection of variables, which to be generic I'll write $X_1, X_2, \ldots X_p$. These may be discrete, continuous, or even vectors; it doesn't matter. We represent these visually as nodes in a graph. There are arrows connecting some of these nodes. If an

arrow runs from X_i to X_j, then X_i is a **parent** of X_j. This is, as the name "parent" suggests, an anti-symmetric relationship, i.e., X_j cannot also be the parent of X_i. This is why we use an arrow, and why the graph is **directed**[1]. We write the set of all parents of X_j as parents(j); this generalizes the notion of the factors which an observable loads on to. The joint distribution "decomposes according to the graph":

$$p(X_1, X_2, \ldots X_p) = \prod_{i=1}^{p} p(X_i | X_{\text{parents}(i)}) \tag{21.4}$$

If X_i has no parents, because it has no incoming arrows, take $p(X_i|X_{\text{parents}(i)})$ just to be the marginal distribution $p(X_i)$. Such variables are called **exogenous**; the others, with parents, are **endogenous**. An unfortunate situation could arise where X_1 is the parent of X_2, which is the parent of X_3, which is the parent of X_1. Perhaps, under some circumstances, we could make sense of this and actually calculate with Eq. 21.4, but the general practice is to rule it out by assuming the graph is **acyclic**, i.e., that it has no cycles, i.e., that we cannot, by following a series of arrows in the graph, go from one node to other nodes and ultimately back to our starting point. Altogether we say that we have a **directed acyclic graph**, or **DAG**, which represents the direct dependencies between variables.[2]

What good is this? The primary virtue is that if we are dealing with a DAG model, the graph tells us all the dependencies we need to know; those are the conditional distributions of variables on their parents, appearing in the product on the right hand side of Eq. 21.4. (This includes the distribution of the exogeneous variables.) This fact has two powerful sets of implications, for probabilistic reasoning and for statistical inference.

Let's take inference first, because it's more obvious: all that we have to estimate are the conditional distributions $p(X_i|X_{\text{parents}(i)})$. We do not have to estimate the distribution of X_i given *all* of the other variables, unless of course they are all parents of X_i. Since estimating distributions, or even just regressions, conditional on many variables is hard, it is extremely helpful to be able to read off from the graph which variables we can *ignore*. Indeed, if the graph tells us that X_i is exogeneous, we don't have to estimate it conditional on anything, we just have to estimate its marginal distribution.

21.2.1 Conditional Independence and the Markov Property

The probabilistic implication of Eq. 21.4 is perhaps even more important, and that has to do with conditional independence. Pick any two variables X_i and X_j, where X_j is not a parent of X_i. Consider the distribution of X_i conditional on its parents *and* X_j. There are two possibilities. (i) X_j is not a descendant of X_i. Then we can see that X_i and X_j are conditionally independent. This is true *no matter what* the actual conditional distribution functions involved are; it's just implied by the joint

[1] See Appendix E for a brief review of the ideas and jargon of graph theory.

[2] See §21.4 for remarks on undirected graphical models, and graphs with cycles.

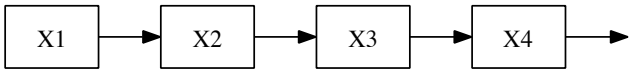

Figure 21.2: DAG for a discrete-time Markov process. At each time t, X_t is the child of X_{t-1} and the parent of X_{t+1}.

distribution respecting the graph. (ii) Alternatively, X_j is a descendant of X_i. Then in general they are not independent, even conditional on the parents of X_i. So the graph implies that certain conditional independence relations will hold, but that others in general will *not* hold.

As you know from your probability courses, a sequence of random variables $X_1, X_2, X_3, \ldots$ forms a **Markov process**[3] when "the past is independent of the future given the present": that is,

$$X_{t+1} \perp\!\!\!\perp (X_{t-1}, X_{t-2}, \ldots X_1) | X_t \tag{21.5}$$

from which it follows that

$$(X_{t+1}, X_{t+2}, X_{t+3}, \ldots) \perp\!\!\!\perp (X_{t-1}, X_{t-2}, \ldots X_1) | X_t \tag{21.6}$$

which is called the **Markov property**. DAG models have a similar property: if we take any collection of nodes I, it is independent of its non-descendants, given its parents:

$$X_I \perp\!\!\!\perp X_{\text{non-descendants}(I)} | X_{\text{parents}(I)} \tag{21.7}$$

This is the **directed graph Markov property**. The ordinary Markov property is in act a special case of this, when the graph looks like Figure 21.2[4].

21.3 Examples of DAG Models and Their Uses

Factor models are examples of DAG models (as we've seen). So are mixture models (Figure 21.3) and Markov chains (see above). DAG models are considerably more flexible, however, and can combine observed and unobserved variables in many ways.

Consider, for instance, Figure 21.4. Here there are two exogeneous variables, labeled "Smoking" and "Asbestos". Everything else is endogenous. Notice that "Yellow teeth" is a child of "Smoking" alone. This does not mean that (in the model)

[3] After the Russian mathematician A. A. Markov, who introduced the theory of Markov processes in the course of a mathematical dispute with his arch-nemesis, to show that probability and statistics could apply to dependent events, and hence that Christianity was not *necessarily* true (I am not making this up: Basharin *et al.*, 2004).

[4] To see this, take the "future" nodes, indexed by $t+1$ and up, as the set I. Their parent consists just of X_t, and all their non-descendants are the even earlier nodes at times $t-1$, $t-2$, etc.

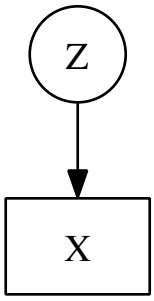

Figure 21.3: DAG for a mixture model. The latent class Z is exogenous, and the parent of the observable random vector X. (If the components of X are conditionally independent given Z, they could be represented as separate boxes on the lower level.

whether someone's teeth get yellowed (and, if so, how much) is a function of smoking alone; it means that whatever other influences go into that are independent of the rest of the model, and so unsystematic that we can think about those influences, taken together, as noise.

Continuing, the idea is that how much someone smokes influences how yellow their teeth become, and also how much tar builds up in their lungs. Tar in the lungs, in turn, leads to cancer, as does by exposure to asbestos.

Now notice that, in this model, teeth-yellowing will be unconditionally dependent on, i.e., associated with, the level of tar in the lungs, because they share a common parent, namely smoking. Yellow teeth and tarry lungs will however be conditionally independent given that parent, so if we control for smoking we should not be able to predict the state of someone's teeth from the state of their lungs or vice versa.

On the other hand, smoking and exposure to asbestos are independent, at least in this model, as they are both exogenous[5]. Conditional on whether someone has cancer, however, smoking and asbestos will become *dependent*.

To understand the logic of this, suppose (what is in fact true) that both how much someone smokes and how much they are exposed to asbestos raises the risk of cancer. Conditional on not having cancer, then, one was probably exposed to little of either tobacco smoke or asbestos. Conditional on both not having cancer and having

[5]If we had two variables which in some physical sense were exogenous but dependent on each other, we would represent them in a DAG model by either a single *vector-valued* random variable (which would get only one node), or as children of a latent unobserved variable, which was truly exogenous.

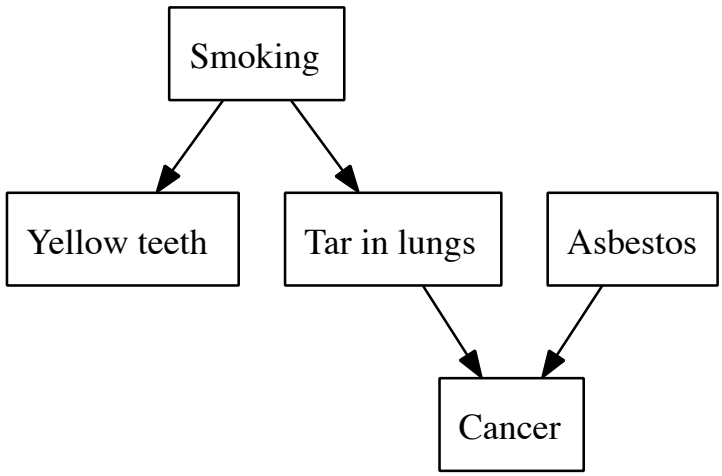

Figure 21.4: DAG model indicating (hypothetical) relationships between smoking, asbestos, cancer, and covariates.

been exposed to a high level of asbestos, one probably was exposed to an unusually low level of tobacco smoke. Vice versa, no cancer plus high levels of tobacco tend to imply especially little exposure to asbestos. We thus have created a negative association between smoking and asbestos by conditioning on cancer. Naively, a regression where we "controlled for" cancer would in fact tell us that exposure to asbestos keeps tar from building up in the lungs, prevents smoking, and whitens teeth.

More generally, conditioning on a third variable can *create* dependence between otherwise independent variables, when what we are conditioning on is a common *descendant* of the variables in question.[6] This conditional dependence is *not* some kind of finite-sample artifact or error — it really is there in the joint probability distribution. If all we care about is prediction, then it is perfectly legitimate to use it. In the world of Figure 21.4, it really is true that you can predict the color of someone's teeth from whether they have cancer and how much asbestos they've been exposed to, so if that's what you want to predict[7], why not use that information? But if you want to do more than just make predictions without understanding, if you want to understand the structure tying together these variables, if you want to do *science*, if you don't want to go around telling yourself that asbestos whitens teeth, you really do need to know the graph.

21.3.1 Missing Variables

Suppose that we do cannot observe one of the variables, such as the quantity of tar in the lungs, but we somehow know all of the conditional distributions required by the graph. (Tar build-up in the lungs might indeed be hard to measure for living people.) Because we have a joint distribution for *all* the variables, we could estimate the conditional distribution of one of them given the rest, using the definition of conditional probability and of integration:

$$p(X_i|X_1,X_2,X_{i-1},X_{i+1},X_p) = \frac{p(X_1,X_2,X_{i-1},X_i,X_{i+1},X_p)}{\int p(X_1,X_2,X_{i-1},x_i,X_{i+1},X_p)dx_i} \tag{21.8}$$

We could in principle do this for *any* joint distribution. When the joint distribution comes from a DAG model, however, we can simplify this considerably. Recall that, from Eq. 21.7, X_i conditioning on its parents makes X_i independent of all its non-descendants. We can therefore drop from the conditioning everything which isn't either a parent of X_i, or a descendant. In fact, it's not hard to see that given the children of X_i, its more remote descendants are also redundant. Actually *doing* the calculation then boils down to a version of the EM algorithm.[8]

[6] Economists, psychologists, and other non-statisticians often repeat the advice that if you want to know the effect of X on Y, you should not condition on Z when Z is endogenous. This is bit of folklore is an incorrect relic of the days of ignorance, though it shows a sound indistinct groping towards a truth those people were unable to grasp. If we want to know whether asbestos is associated with tar in the lungs, conditioning on the yellowness of teeth is fine, even though that is an endogenous variable.

[7] Maybe you want to guess who'd be interested in buying whitening toothpaste.

[8] Graphical models, especially directed ones, are often called "Bayes nets" or "Bayesian networks", because this equation is, or can be seen as, a version of Bayes's rule. Since of course it follows directly from the definition of conditional probability, there is really nothing Bayesian about them.

If we observe only a subset of the other variables, we can still use the DAG to determine which ones actually matter to estimating X_i, and which ones are superfluous. The calculations then however become much more intricate.[9]

21.4 Non-DAG Graphical Models: Undirected Graphs and Directed Graphs with Cycles

This section is optional, as, for various reasons, we will not use these models in this course.

21.4.1 Undirected Graphs

There is a lot of work on probability models which are based on *undirected* graphs, in which the relationship between random variables linked by edges is completely symmetric, unlike the case of DAGs[10]. Since the relationship is symmetric, the preferred metaphor is not "parent and child", but "neighbors". The models are sometimes called **Markov networks** or **Markov random fields**, but since DAG models have a Markov property of their own, this is not a happy choice of name, and I'll just call them "undirected graphical models".

The key Markov property for undirected graphical models is that any set of nodes I is independent of the rest of the graph given its neighbors:

$$X_I \perp\!\!\!\perp X_{\text{non-neighbors}(I)} | X_{\text{neighbors}(I)} \tag{21.9}$$

This corresponds to a factorization of the joint distribution, but a more complex one than that of Eq. 21.4, because a symmetric neighbor-of relation gives us no way of *ordering* the variables, and conditioning the later ones on the earlier ones. The trick turns out to go as follows. First, as a bit of graph theory, a **clique** is a set of nodes which are all neighbors of each other, and which cannot be expanded without losing that property. We write the collection of all cliques in a graph G as cliques(G). Second, we introduce **potential functions** ψ_c which take clique configurations and return non-negative numbers. Third, we say that a joint distribution is a **Gibbs distribution**[11] when

$$p(X_1, X_2, \ldots X_p) \propto \prod_{c \in \text{cliques}(G)} \psi_c(X_{i \in c}) \tag{21.10}$$

That is, the joint distribution is a product of factors, one factor for each clique. Frequently, one introduces what are called **potential functions**, $U_c = \log \psi_c$, and then one has

$$p(X_1, X_2, \ldots X_p) \propto e^{-\sum_{c \in \text{cliques}(G)} U_i(X_{i \in c})} \tag{21.11}$$

[9] There is an extensive discussion of relevant methods in Jordan (1998).

[10] I am told that this is more like the idea of causation in Buddhism, as something like "co-dependent origination", than the asymmetric one which Europe and the Islamic world inherited from the Greeks (especially Aristotle), but you would really have to ask a philosopher about that.

[11] After the American physicist and chemist J. W. Gibbs, who introduced such distributions as part of **statistical mechanics**, the theory of the large-scale patterns produced by huge numbers of small-scale interactions.

The key correspondence is what is sometimes called the **Gibbs-Markov theorem**: a distribution is a Gibbs distribution with respect to a graph G if, and only if, it obeys the Markov property with neighbors defined according to G.[12].

In many practical situations, one combines the assumption of an undirected graphical model with the further assumption that the joint distribution of all the random variables is a multivariate Gaussian, giving a **Gaussian graphical model**. An important consequence of this assumption is that the graph can be "read off" from the inverse of the covariance matrix Σ, sometimes called the **precision matrix**. Specifically, there is an edge linking X_i to X_j if and only if $(\Sigma^{-1})_{ij} \neq 0$. (See Lauritzen (1996) for an extensive discussion.) These ideas sometimes still work for non-Gaussian distributions, when there is a natural way of transforming them to be Gaussian (Liu *et al.*, 2009), though it is unclear just how far that goes.

Further reading Markov random fields where the graph is a regular lattice are used extensively in spatial statistics. Good introductory-level treatments are provided by Kindermann and Snell (1980) (the full text of which is free online), and by Guttorp (1995), which also covers the associated statistical methods. Winkler (1995) is also good, but presumes more background in statistical theory. (I would recommend reading it after Guttorp.) Guyon (1995) is at a similar level of sophistication to Winkler, but, unlike the previous references, emphasizes the situations where the graph is *not* a regular lattice. Griffeath (1976), while presuming more probability theory on the part of the reader, is extremely clear and insightful, including what is simultaneously one of the deepest and most transparent proofs of the Gibbs-Markov theorem. Lauritzen (1996) is a mathematically rigorous treatment of graphical models from the viewpoint of theoretical statistics, covering both the directed and undirected cases.

If you are curious about Gibbs distributions in, so to speak, their natural habitat, the book by Sethna (2006), also free online, is the best introduction to statistical mechanics I have seen, and presumes very little knowledge of actual physics on the part of the reader. Honerkamp (2002) is less friendly, but tries harder to make connections to statistics. If you already know what an exponential family is, then Eq. 21.11 is probably extremely suggestive, and you should read Mandelbrot (1962).

21.4.2 Directed but Cyclic Graphs

Much less work has been done on directed graphs with cycles. It is very hard to give these a causal interpretation, in the fashion described in the next chapter. Feedback processes are of course very common in nature and technology, and one might think to represent these as cycles in a graph. A model of a thermostat, for instance, might

[12] This theorem was proved, in slightly different versions, under slightly different conditions, and by very different methods, more or less simultaneously by (alphabetically) Dobrushin, Griffeath, Grimmett, and Hammersley and Clifford, and almost proven by Ruelle. In the statistics literature, it has come to be called the "Hammersley-Clifford" theorem, for no particularly good reason. In my opinion, the clearest and most interesting version of the theorem is that of Griffeath (1976), an elementary exposition of which is given by Pollard (`http://www.stat.yale.edu/~pollard/Courses/251.spring04/Handouts/Hammersley-Clifford.pdf`). (Of course, Griffeath was one of my Ph.D. supervisors, so discount accordingly.) Calling it the "Gibbs-Markov theorem" says more about the content, and is fairer to all concerned.

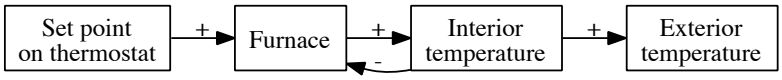

Figure 21.5: Directed but cyclic graphical model of a feedback loop. Signs (+, − on arrows are "guides to the mind". Cf. Figure 21.6.

have variables for the set-point temperature, the temperature outside, how much the furnace runs, and the actual temperature inside, with a cycle between the latter two (Figure 21.5).

Thinking in this way is however simply sloppy. It always takes *some* time to traverse a feedback loop, and so the cycle really "unrolls" into an acyclic graph linking similar variables at *different* times (Figure 21.6). Sometimes[13], it is clear that when people draw a diagram like Figure 21.5, the incoming arrows really refer to the change, or rate of change, of the variable in question, so it is merely a visual short-hand for something like Figure 21.6.

Directed graphs with cycles are thus primarily useful when measurements are so slow or otherwise imprecise that feedback loops cannot be unrolled into the actual dynamical processes which implement them, and one is forced to hope that one can reason about equilibria instead[14]. If you insist on dealing with cyclic directed graphical models, see Richardson (1996); Lacerda *et al.* (2008) and references therein.

21.5 Further Reading

The paper collection Jordan (1998) is actually extremely good, unlike most collections of edited papers; Jordan and Sejnowski (2001) is also useful. Lauritzen (1996) is thorough but more mathematically demanding. The books by Spirtes *et al.* (1993, 2001) and by Pearl (1988, 2000, 2009b) are classics, especially for their treatment of causality, of which much more soon. Glymour (2001) discusses applications to psychology.

While I have presented DAG models as an outgrowth of factor analysis, their historical ancestry is actually closer to the "path analysis" models introduced by the great mathematical biologist Sewall Wright in the 1920s to analyze processes of development and genetics. These proved extremely influential in psychology. Loehlin (1992) is user-friendly, though aimed at psychologists with less mathematical sophistication than students taking this course. Li (1975), while older, is very enthusiastic and has many interesting applications.

[13] As in Puccia and Levins (1985), and the `LoopAnalyst` package based on it (Dinno, 2009).

[14] Economists are fond of doing so, generally without providing any rationale, based in economic theory, for supposing that equilibrium *is* a good approximation.

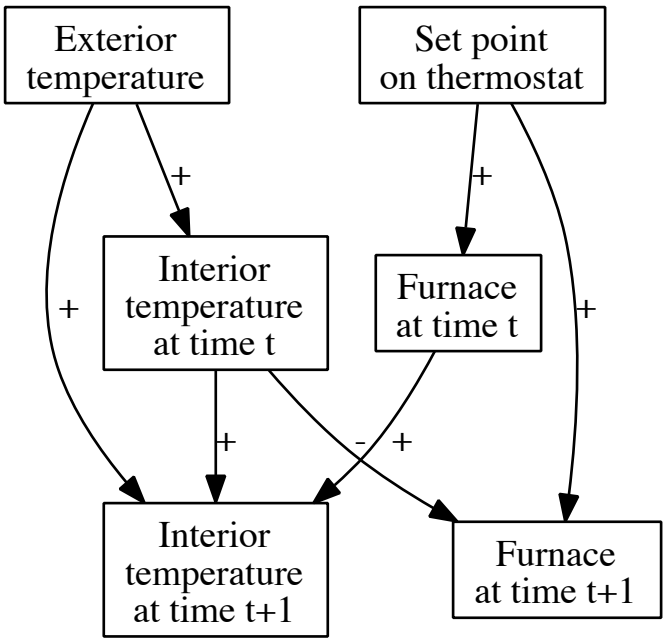

Figure 21.6: Directed, acyclic graph for the situation in Figure 21.5, taking into account the fact that it takes time to traverse a feedback loop. One should imagine this repeating to times $t+2$, $t+3$, etc., and extending backwards to times $t-1$, $t-2$, etc., as well. Notice that there are no longer any cycles.