2
Statistical modelling

2.1 Introduction

A statistical model for a set of random quantities X is a family of probability distributions for X, usually represented, by me, as

$$\{\mathcal{X}, \Omega, f_X\}$$

where $\mathcal{X}$ is the realm of X, Ω is the parameter space, and f_X is a set of probability mass functions (PMFs) indexed by $\theta \in \Omega$. This is the 'old school' definition of a statistical model, which is still the prevailing definition in the Frequentist approach to inference, and also in the 'neo-Bayesian' approach to inference, in which the statistical model is augmented with a prior distribution π.

The modern Bayesian approach has a different concept of a statistical model, but which, confusingly, retains the same notation and terminology. In the modern Bayesian approach a statistical model for X is a single probability distribution, but expressed for (Θ, X) rather than just X, where Θ are additional random variables whose presence simplifies the expression of the joint distribution $f_{\Theta,X}$. The modern Bayesian approach is consistent with the neo-Bayesian approach, if we write

$$f_{\Theta,X}(\theta, x) = \pi(\theta) \cdot f_X(x; \theta).$$

The innovation in the modern Bayesian approach is that it transcends the rigid distinction between random quantities and parameters that is the starting-point for the Frequentist and Neo-Bayesian approaches, with a much more liberal interpretation of 'parameter'. Here is the basic maxim of the modern Bayesian approach:

Definition 2.1 (Maxim of statistical modelling)**.** Obtain the desired marginal distribution f_X by introducing additional random variables Θ, specifying the joint distribution $f_{\Theta,X}$ using simplifications from conditional independence, and then marginalizing over Θ.

This maxim only works because of the simplifying properties of conditional independence assumptions; otherwise it would be harder to specify $f_{\Theta,X}$ than f_X. So the basic guidance when considering what Θ to introduce is whether these Θ can reasonably induce lots of conditional independence, as discussed below.

Furthermore, this maxim requires that we can marginalize over possibly large collections of random variables. Thirty years ago this was not possible, but improvements in computer power and in computing algorithms have changed that. In particular, Markov Chain Monte Carlo (MCMC) has 'solved' the problem of marginalization for many of the common statistical models.

2.2 *Notation*

It is not usual to have a whole section on notation, but representing modern statistical models is a thorny topic.

In general, X is a collection of random quantities, and Θ is a collection of *random variables*. I write 'random variables' because I want to maintain the notion that X are operationally defined. But Θ are introduced by us, at our convenience, and may not be operationally defined; hence they need a different name. For convenience I will treat $\mathcal{X}$, the realm of X, as finite and Ω, the realm of Θ, as uncountable. This is also the realistic case. But it means that probability statements are complicated, because $X = x$ can have a non-zero probability, but $\Theta = \theta$ cannot. I come back to this below.

Let $X = (X_1, \dots, X_m)$ be a collection of m random quantities. If $A = (a_1, \dots, a_k)$ where the a_j are distinct elements in $\{1, \dots, m\}$, then $X_A := (X_{a_1}, \dots, X_{a_k})$. It is convenient to write $X_{i:j}$ when $A = (i, \dots, j)$. Exactly the same conventions apply to $\mathcal{X}$ and to x.

Let f_X be a PMF of X, i.e.

$$f_X(x) = \Pr(X_1 = x_1 \wedge \cdots \wedge X_m = x_m),$$

and '$\wedge$' denotes 'and'.[1] Then the marginal PMF of X_A (i.e. marginalizing over X_B) is

$$f_{X_A}(x_A) = \sum_{x_B \in \mathcal{X}_B} f_X(x_A, x_B)$$

where B is the complement of A in $\{1, \dots, m\}$, and there is no ambiguity about how to combine x_A and x_B into an element of $\mathcal{X}$; hence we just write 'x_A, x_B' as the argument to f_X.

There is a more parsimonious notation for a PMF, in which the identity of the random quantities is inferred from the arguments, taking advantage of the convention that x_A is a arbitrary value in $\mathcal{X}_A$, the realm of X_A. Thus we write

$$\mathrm{p}(x_A) := f_{X_A}(x_A). \tag{2.1}$$

Less ink usually means more clarity, so I will use this convention wherever I can.

The attraction of the 'p' notation is that is can be extended to Θ, partly concealing the difficulty that the realm of Θ, denoted Ω, is uncountable. By convention, if Ω is uncountable, then

$$\mathrm{p}(\theta) := \Pr\{\Theta \in [\theta, \theta + \mathrm{d}\theta)\} = \pi(\theta) \cdot \mathrm{d}\theta, \tag{2.2}$$

[1] This is the standard symbol for conjunction, from the propositional calculus.

the second equality following in the case where Θ is treated as absolutely continuous, for which uncertainty about Θ is represented by a probability density function (PDF), denoted π. Therefore we can write hybrid probabilities such as $\mathrm{p}(\theta, x)$ and these will follow the usual rules of the probability calculus. For example,

$$\begin{aligned} \mathrm{p}(\theta, x) &= \Pr\{\Theta \in [\theta, \theta + \mathrm{d}\theta) \wedge X = x\} \\ &= \Pr\{\Theta \in [\theta, \theta + \mathrm{d}\theta)\} \cdot \Pr\{X = x \mid \Theta \in [\theta, \theta + \mathrm{d}\theta)\} \\ &= \mathrm{p}(\theta) \cdot \mathrm{p}(x \mid \theta). \end{aligned}$$

Here I have used the defining property of a conditional probability, which will be explained in Section 2.3.

2.3 *Conditional probabilities*

In this section, $A, B, \ldots$ will be arbitrary propositions, i.e. statements which are either FALSE or TRUE. They can be combined into new propositions using the *propositional calculus*. The axioms of probability provide a framework for reasoning about the truths of a set of propositions in the presence of uncertainty. These axioms can be used to prove simple but important results such as

$$\text{If } \Pr(B) = 0, \text{ then } \Pr(A \wedge B) = 0,$$

which I will use several times below.

I define *conditional probabilities* implicitly. $\Pr(A \mid B)$ is the conditional probability of A given B exactly when

$$\Pr(A \wedge B) = \Pr(A \mid B) \cdot \Pr(B). \tag{2.3}$$

Under this definition, $\Pr(A \mid B)$ is arbitrary (between 0 and 1) when $\Pr(B) = 0$, since in this case (2.3) reads $0 = \Pr(A \mid B) \cdot 0$; otherwise $\Pr(A \mid B)$ is the unique value

$$\Pr(A \mid B) = \frac{\Pr(A \wedge B)}{\Pr(B)}, \quad \Pr(B) > 0. \tag{2.4}$$

It is easily checked that if $\Pr(B) > 0$, then $\Pr(\bullet \mid B)$ obeys the axioms of probability. The following result is crucial in extending conditional probabilities to more complex situations.

Theorem 2.1 (Extension theorem)**.** *If* $\Pr(C) > 0$*, then*

$$\Pr(A \wedge B \mid C) = \Pr(A \mid C) \cdot \Pr(B \mid A \wedge C). \tag{2.5}$$

Proof. First, suppose that $\Pr(A \wedge C) = 0$. In this case $\Pr(A \mid C) = 0$ by definition, because $\Pr(C) > 0$; and $\Pr(A \wedge B \mid C) = 0$, because $\Pr(\bullet \mid C)$ obeys the axioms of probability. Hence (2.5) has the form $0 = 0 \cdot \Pr(B \mid A \wedge C)$, and the result holds in this case, with the value of $\Pr(B \mid A \wedge C)$ being arbitrary.

Now suppose that $\Pr(A \wedge C) > 0$. In this case,

$$\begin{aligned} \Pr(A \wedge B \wedge C) &= \Pr(A \wedge C) \cdot \frac{\Pr(A \wedge B \wedge C)}{\Pr(A \wedge C)} \\ &= \Pr(A \wedge C) \cdot \Pr(B \mid A \wedge C). \end{aligned}$$

Now divide through by $\Pr(C)$, which is non-zero, to complete the proof. □

The Extension theorem can be iterated to provide a factorization of any conjunction.[2]

Theorem 2.2 (Telescope theorem)**.** *Let $A_1, \dots, A_m, C$ be a set of propositions, and write $A_{i:j} := A_i \wedge \cdots \wedge A_j$. If $\Pr(C) > 0$, then*

$$\Pr(A_{1:i} \mid C) = \Pr(A_1 \mid C) \cdot \prod_{j=2}^{i} \Pr(A_j \mid A_{1:(j-1)} \wedge C). \tag{2.6}$$

Proof. In Theorem 2.1, set $A \leftarrow A_1$, $B \leftarrow A_{2:i}$, and $C \leftarrow C$, to give

$$\Pr(A_{1:i} \mid C) = \Pr(A_1 \wedge A_{2:i} \mid C) = \Pr(A_1 \mid C) \cdot \Pr(A_{2:i} \mid A_1 \wedge C).$$

Now expand out the second term on the righthand side, by repeated use of Theorem 2.1. If for some j, $\Pr(A_{1:(j-1)} \wedge C) = 0$, then the additional conditional probabilities have arbitrary values, but the result, which has the form $0 = 0 \cdot \Pr(A_{j:i} \mid A_{1:(j-1)} \wedge C)$, still holds. □

2.4 *Conditional independence*

In this section I will write 'f_X' or 'f_{X_A}' when I need to refer explicitly to the joint PMF of X or the marginal PMF of X_A, but I will use the 'p' notation in expressions, for clarity.

The Telescope theorem (Theorem 2.2) asserts that if $X = (X_1, \dots, X_m)$, then $f_X(x)$ equals

$$\mathrm{p}(x_1, \dots, x_m) = \mathrm{p}(x_1) \cdot \prod_{j=2}^{m} \mathrm{p}(x_j \mid x_{1:(j-1)}). \tag{2.7}$$

The Telescope theorem suggest that we can construct the joint PMF f_X by thinking conditionally, one element (or one block of elements) at a time, and them multiplying them all together.

Of course this is still a lot of work. $f_{X_j \mid X_{1:(j-1)}}$ is a function with j arguments, which has to be a PMF for X_j for every $x_{1:(j-1)} \in \mathcal{X}_{1:(j-1)}$. Without some simplifications, this conditional approach to specifying f_X is unlikely to be easier than specifying f_X directly. Conditional independence is the crucial simplification. Here is the formal definition.

Definition 2.2 (Conditional independence)**.** Let A, B, C be disjoint subsets of $\{1, \dots, m\}$. X_A is conditionally independent of X_B given X_C exactly when

$$\mathrm{p}(x_A, x_B \mid x_C) = \mathrm{p}(x_A \mid x_C) \cdot \mathrm{p}(x_B \mid x_C) \tag{2.8}$$

whenever $\mathrm{p}(x_C) > 0$. In this case we write

$$X_A \perp\!\!\!\perp X_B \mid X_C.$$

[2] I'm calling this the 'Telescope theorem' because I like the image, but elsewhere you may see it called the 'chain rule for probabilities'.

The practical implications of condional independence are captured in the following result.

Theorem 2.3. *The following statements are equivalent:*

1. $X_A \perp\!\!\!\perp X_B \mid X_C$.

2. *If* $\mathrm{p}(x_B, x_C) > 0$, *then*

$$\mathrm{p}(x_A \mid x_B, x_C) = \mathrm{p}(x_A \mid x_C). \tag{2.9}$$

Proof.

(1) $\implies$ (2). If $\mathrm{p}(x_B, x_C) > 0$, then $\mathrm{p}(x_B \mid x_C) > 0$. Dividing (2.8) by $\mathrm{p}(x_B \mid x_C)$ gives (2.9), after applying the Extension theorem (Theorem 2.1).

(2) $\implies$ (1). If $\mathrm{p}(x_C) > 0$ then

$$\mathrm{p}(x_A, x_B \mid x_C) = \mathrm{p}(x_A \mid x_B, x_C) \cdot \mathrm{p}(x_B \mid x_C) \tag{\dagger}$$

by Theorem 2.1. If $\mathrm{p}(x_B, x_C) = 0$ then $\mathrm{p}(x_A \mid x_B, x_C)$ is arbitrary, and we can set it equal to $\mathrm{p}(x_A \mid x_C)$ in (†), to give (2.8). If $\mathrm{p}(x_B, x_C) > 0$, then we can substitute from (2.9) into (†) to give (2.8), as required.

□

Eq. (2.9) states that if $X_A \perp\!\!\!\perp X_B \mid X_C$, then knowledge of X_B is irrelevant when predicting X_A with knowledge of X_C. What is deeply mysterious is this irrelevance relationship is symmetric, i.e.

$$X_A \perp\!\!\!\perp X_B \mid X_C \iff X_B \perp\!\!\!\perp X_A \mid X_C, \tag{2.10}$$

which follows from the symmetry of the original definition, in (2.8). Dawid (1998) discusses this and other properties of conditional independence.

Here is an important implication of conditional independence:

$$X_A \perp\!\!\!\perp X_B \mid X_C \implies X_A \perp\!\!\!\perp g(X_B) \mid X_C \tag{2.11}$$

for all g. This includes the special case where $g(x_B)$ is a subset of the elements of X_B. The proof is straightforward. By symmetry, X_A is irrelevant when predicting X_B in the presence of X_C. Therefore X_A must be irrelevant in predicting any function of X_B in the presence of X_C, and so $X_A \perp\!\!\!\perp g(X_B) \mid X_C$.

One special case of conditional independence comes up frequently when modelling. If

$$X_A \perp\!\!\!\perp X_B \mid Y \tag{2.12}$$

for all possible disjoint sets A and B, then X is *mutually conditionally independent given* Y. I write this as

$$\models X \mid Y. \tag{2.13}$$

It is straightforward to prove that

$$\vDash X \mid Y \iff f_{X|Y}(x \mid y) = \prod_{i=1}^{m} f_{X_i|Y}(x_i \mid y). \qquad (2.14)$$

Mutual conditional independence is a very powerful modelling assumption, because it factorizes a conditional PMF with $m + 1$ arguments into the product of m conditional PMFs each with only two arguments. It is common to have the additional simplification that $f_{X_i|Y}$ is the same for all i. In this case we would say that X is *mutually conditionally independent and identically distributed (IID) given* Y.

* * *

For completeness, I also mention 'unconditional' independence, in which there is nothing to condition on. X_A and X_B are *independent* exactly when

$$\mathrm{p}(x_A, x_B) = \mathrm{p}(x_A) \cdot \mathrm{p}(x_B), \qquad (2.15)$$

which is written $X_A \perp\!\!\!\perp X_B$. Likewise, X can be mutually independent, which I write as $\vDash X$, and, as a special case of this, (mutually) IID.

Here are two very important things to remember about the non-relationship between independence and conditional independence:

1. $X \perp\!\!\!\perp Y \;\not\Longrightarrow\; X \perp\!\!\!\perp Y \mid Z$. For example, X and Y might be the values from two independent rolls of a dice, and Z might be the sum of the two values.
2. $X \perp\!\!\!\perp Y \mid Z \;\not\Longrightarrow\; X \perp\!\!\!\perp Y$. For example, X might be electricity generated in a hydro-electric plant, Z might be depth of water behind the dam, and Y might be recent rainfall on the catchment above the dam.

These two non-relationships can be proved probabilistically, but the examples are more vivid. Independence and conditional independence are two different things.

2.5 Graphical models

When we construct f_X using conditional independence, we use (2.9) to simplify the terms in the Telescope theorem. In particular, we order the X_i's so that when we write out the Telescope factorization, only a subset of $X_{1:(j-1)}$ are relevant in $f_{X_j|X_{1:(j-1)}}$. Let this subset be denoted $\mathrm{pa}_j \subseteq \{1, \dots, j-1\}$, where '$\mathrm{pa}_j$' is read as 'parents of X_j'. So the Telescope theorem is also written

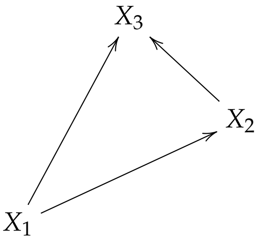
$$\mathrm{p}(x) = \mathrm{p}(x_1) \cdot \prod_{j=2}^{m} \mathrm{p}(x_j \mid x_{\mathrm{pa}_j}). \qquad (2.16)$$

According to Theorem 2.3,

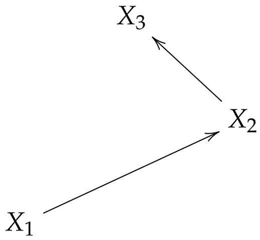
$$X_j \perp\!\!\!\perp X_{\overline{\mathrm{pa}}_j} \mid X_{\mathrm{pa}_j} \quad j = 2, \dots, m, \qquad (2.17)$$

where $\overline{\mathrm{pa}}_j$ is the complement of pa_j in $\{1, \ldots, j-1\}$. So the set $\{\mathrm{pa}_2, \ldots, \mathrm{pa}_m\}$ represents our conditional independence modelling for X. If pa_j is a strict subset of $\{1, \ldots, j-1\}$ then conditional independence modelling has simplified the task of specifying $f_{X_j \mid X_{1:(j-1)}}$ by reducing the number of arguments this function requires.

There is a simple way to visualize the set of pa_j's, as a *directed acyclic graph (DAG)*. In a DAG each X_j is a vertex (or node) and there is an edge from X_i from X_j exactly when $i \in \mathrm{pa}_j$, or X_i is a 'parent' of X_j. Here are two DAGS for $m = 3$:



and



In the lefthand case there is no conditional independence, the set of pa's is $\{\mathrm{pa}_2 = \{1\}, \mathrm{pa}_3 = \{1, 2\}\}$, and

$$\mathrm{p}(x_1, x_2, x_3) = \mathrm{p}(x_1) \cdot \mathrm{p}(x_2 \mid x_1) \cdot \mathrm{p}(x_3 \mid x_1, x_2).$$

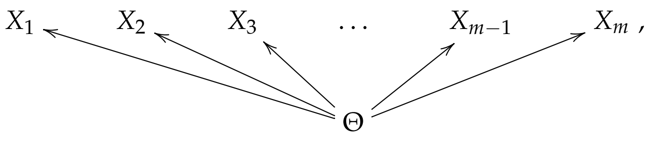
In the righthand case the edge from X_1 to X_3 is missing, and the set of pa's is $\{\mathrm{pa}_2 = \{1\}, \mathrm{pa}_3 = \{2\}\}$. So in the righthand case we have $X_3 \perp\!\!\!\perp X_1 \mid X_2$, and

$$\mathrm{p}(x_1, x_2, x_3) = \mathrm{p}(x_1) \cdot \mathrm{p}(x_2 \mid x_1) \cdot \mathrm{p}(x_3 \mid x_2).$$

Some stochastic processes are actually defined by their conditional independencies, or, equivalently, their DAGs. If X is a Markov process, then its DAG is

$$X_1 \longrightarrow X_2 \longrightarrow X_3 \qquad \ldots \qquad X_{m-1} \longrightarrow X_m \ .$$

If X is mutually conditionally independent given Θ (i.e. $\models X \mid \Theta$) then the DAG of (Θ, X) is



(2.18)

from (2.14).

DAGs can be used to understand the maxim of statistical modelling (Definition 2.1), given in Section 2.1. Consider, for example, (2.18). Suppose that we are interested in the marginal distribution of X, where the X's are a set of random quantities which are similar but not identical. We introduce Θ to simplify the joint distribution of (Θ, X), by asserting that $\models X \mid \Theta$. In this case, we need to specify f_Θ and $f_{X_j \mid \Theta}$ for $j = 1, \ldots, m$. From (2.14),

$$\mathrm{p}(\theta, x_{1:i}) = \mathrm{p}(\theta) \cdot \mathrm{p}(x_{1:i} \mid \theta) = \mathrm{p}(\theta) \cdot \prod_{j=1}^{i} \mathrm{p}(x_j \mid \theta). \qquad (2.19)$$

Now marginalize over Θ to give

$$\mathrm{p}(x_{1:i}) = \int_{\Omega} \mathrm{p}(\theta) \cdot \prod_{j=1}^{i} \mathrm{p}(x_j \mid \theta)\, \mathrm{d}\theta. \tag{2.20}$$

If we then divide through by $\mathrm{p}(x_{1:(i-1)})$ we see that $\mathrm{p}(x_i \mid x_{1:(i-1)})$ depends on all of the values $x_{1:(i-1)}$. So there is no conditional independence at all in the marginal distribution of X. We start with the simple DAG in (2.18), with only m edges, but when we integrate out Θ we end up with an interesting DAG on X with a full set of $m \cdot (m-1)/2$ edges. In other words, integrating out random variables is a good way to 'complexify' the joint distribution of random quantities.

Here is the general result for what happens to the DAG of X when the random variable Θ is integrated out.

Theorem 2.4. *Let $X = (X_1, \dots, X_m)$ and write $f_{\Theta,X}$ as*

$$\mathrm{p}(\theta, x) = \mathrm{p}(\theta) \cdot \prod_{j=1}^{m} \mathrm{p}(x_j \mid x_{\mathrm{pa}_j}) \tag{2.21}$$

for some random variables Θ, where pa_j are the parents of X_j in $f_{\Theta,X}$, with $0 \in \mathrm{pa}_j$ indicating that Θ is a parent of X_j. Let 'qa_i' denote the parents of X_i in f_X. Then

$$\mathrm{qa}_i = \begin{cases} A_{i-1} \cup \left\{ \bigcup_{j \in A_i} \mathrm{pa}_j \right\} \setminus \{0\} & 0 \in \mathrm{pa}_i \\ \mathrm{pa}_i & 0 \notin \mathrm{pa}_i \end{cases} \tag{2.22}$$

where $A_i = \{ j : 1 \le j \le i$ and $0 \in \mathrm{pa}_j \}$.

Proof. I will write 'pa_j' in place of 'x_{pa_j}', for clarity. Start with the Telescope theorem (Theorem 2.2),

$$\mathrm{p}(\theta, x_{1:i}) = \mathrm{p}(\theta) \cdot \prod_{j=1}^{i} \mathrm{p}(x_j \mid \mathrm{pa}_j). \tag{2.23}$$

Now marginalize over Θ to give

$$\mathrm{p}(x_{1:i}) = \prod_{j \in B_i} \mathrm{p}(x_j \mid \mathrm{pa}_j) \int_{\Omega} \mathrm{p}(\theta) \cdot \prod_{j \in A_i} \mathrm{p}(x_j \mid \mathrm{pa}_j)\, \mathrm{d}\theta, \tag{2.24}$$

where A_i comprises those $j = 1, \dots, i$ for which $0 \in \mathrm{pa}_j$, and B_i the others, for which $0 \notin \mathrm{pa}_j$. Now divide by $\mathrm{p}(x_{1:(i-1)})$ to find $\mathrm{p}(x_i \mid x_{1:(i-1)})$. If $i \in A_i$, then $B_i = B_{i-1}$, and the first term cancels to give

$$\mathrm{p}(x_i \mid x_{1:(i-1)}) = \frac{\int_{\Omega} \mathrm{p}(\theta) \cdot \prod_{j \in A_i} \mathrm{p}(x_j \mid \mathrm{pa}_j)\, \mathrm{d}\theta}{\int_{\Omega} \mathrm{p}(\theta) \cdot \prod_{j \in A_{i-1}} \mathrm{p}(x_j \mid \mathrm{pa}_j)\, \mathrm{d}\theta}.$$

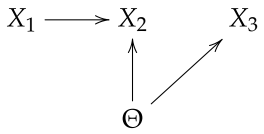
On the other hand, if $i \in B_i$, then $A_i = A_{i-1}$ and the second term cancels, plus most of the first term, to give

$$p(x_i \mid x_{1:(i-1)}) = \mathrm{p}(x_i \mid \mathrm{pa}_i).$$

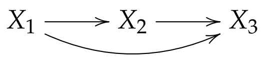
The expression in (2.22) follows directly from these two cases. □

In words, if $\Theta \in \mathrm{pa}_i$ and Θ is marginalized out, then X_i gets an edge from each X_j ($j < i$) for which $\Theta \in \mathrm{pa}_j$, and also an edge from all of the X's that are parents of these X_j's (some of these edges may be duplicates). For example,

from



to



In (2.18), $\mathrm{pa}_j = \{0\}$ and $A_i = \{1, \dots, i\}$. This leads to

$$\mathrm{qa}_i = \{1, \dots, i-1\} \cup \Big\{0 \cup \cdots \cup 0\Big\} \setminus \{0\} = \{1, \dots, i-1\},$$

as claimed. There is a more detailed example in Section 2.6.

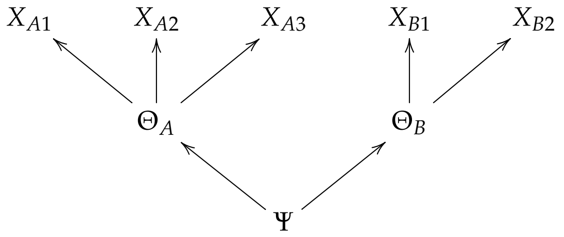
Eq. (2.22) shows that $\mathrm{qa}_i \supset \mathrm{pa}_i \setminus \{0\}$; i.e., when marginalizing over Θ, X_i never loses edges from previous X_j's, and will typically gain edges whenever Θ is a parent of X_i. Hence if Θ is a parent of many X_i's, then marginalizing over Θ can create a DAG with many more edges. We seldom marginalize explicitly, in modern statistical models. Instead, we allow MCMC to do the marginalization for us. The availability of MCMC means we can spend more time thinking about the DAG of (Θ, X), and about the marginal and conditional distributions which appear in (2.16).

2.6 *Hierarchical models*

Hierarchical models are constructed when $X = (X_1, \dots, X_m)$ are similar but not identical. They allow the statistician to control the interrelationship between the X_i's according to known groups (including covariates, in the case of regression). The natural starting point of a hierarchical model is a DAG.

Suppose that $m = 5$, and that (X_1, X_2, X_3) belong in one group, and (X_4, X_5) in another. For example, X_i might be the exam score of a pupil at a school and the two groups might be different classes. We expect X_1 to be more similar to X_2 than X_4. If we had observed (X_1, X_2, X_4) then we would want our prediction for X_3 to be more influenced by x_1^{obs} and x_2^{obs} than x_4^{obs}, and we would want our prediction of X_5 to be more influenced by x_4^{obs} than x_1^{obs} and x_2^{obs}. The beauty of a hierarchical model is that we encode the group structure and then we leave it to the observations to determine the degree to which information from one group affects the prediction of the other.

Write the two groups as $X_A = (X_{A1}, X_{A2}, X_{A3})$ and $X_B = (X_{B1}, X_{B2})$. In general X_{ij} is the jth case in the ith group; this multiple indexing helps with the notation. A common choice of DAG for the illustration is



(2.25)

The X's in group A are similar but not identical, because they share a common parameter, Θ_A; likewise for group B. The two Θ's are similar but not identical, because they share a common parameter Ψ, sometimes termed a *hyperparameter*. This type of modelling can be extended indefinitely, to handle multi-way grouping, and also overlapping group memberships.

The required marginal and conditional distributions can all be read off the DAG. Take (2.25). At the bottom level, we need a marginal distribution for Ψ, f_Ψ. At the middle level we have $\models \Theta \mid \Psi$, which is equivalent to

$$f_{\Theta|\Psi}(\theta \mid \psi) = \prod_{i=A,B} f_{\Theta_i|\Psi}(\theta_i \mid \psi), \tag{2.26}$$

see (2.14). It would be very common to treat $f_{\Theta_i|\Psi}$ as invariant to i, in which case we just need to specify $f_{\Theta_A|\Psi}$, a PDF for Θ_A with parameter Ψ, which we would also use for Θ_B. At the top level we have $\models X_A \mid \Theta_A$ and $\models X_B \mid \Theta_B$. Hence

$$f_{X_i|\Theta_i}(x_i \mid \theta_i) = \prod_{j=1}^{n_A} f_{X_{ij}|\Theta_i}(x_{ij} \mid \theta_i) \quad i = A, B \tag{2.27}$$

($n_A = 3$ and $n_B = 2$). It would be very common to treat $f_{X_{ij}|\Theta_i}$ as invariant to both i and j, in which case we would just need to specify $f_{X_{A1}|\Theta_A}$, a PMF for X_1 with parameter Θ_A. So, in the simplest possible case, we just need to specify

$$f_\Psi, \; f_{\Theta_A|\Psi}, \; \text{and } f_{X_1|\Theta_A}$$

namely one marginal distribution and two conditional distributions. With these three distributions we can handle an arbitrarily large number of groups, and an arbitrarily large number of cases per group.

The DAG alone does not completely specify the statistical model, because although it tells us what distributions we need, it does not tell us the identity of each distribution. In other words, we have to make specific choices for the three distributions given above, which involves being clear about the nature of the parameters and the hyperparameters. Usually, when we do this we find that we need some additional parameters at the top level: typically dispersion parameters, if Θ_i is controlling the expectation of each group. These tend to obscure the structure of the DAG, unfortunately, unless it is constructed using *plates*.[3] Additional parameters can

[3] See `https://en.wikipedia.org/wiki/Plate_notation`. Eq. (2.35) is the complete DAG for (2.28).

be left off the DAG, for clarity, but they need to be expressed in the 'extensive' form of the model, which not only encodes the conditional independence in the DAG, but also the choices for the marginal and joint distributions, and the nature of the parameters.

One choice for the extensive form of the illustration is

$$\begin{aligned} X_{ij} \mid \Theta_i, \sigma^2 &\sim \mathrm{N}(\Theta_i, \sigma^2) && i = A, B;\ j = 1, \dots, n_i && \text{(2.28a)} \\ \Theta_i \mid \mu, \tau^2 &\sim \mathrm{N}(\mu, \tau^2) && i = A, B && \text{(2.28b)} \\ \mu &\sim \mathrm{N}(0, 1000^2) && && \text{(2.28c)} \\ \tau^2 &\sim \mathrm{Ga}(0.001, 0.001) && && \text{(2.28d)} \\ \sigma^2 &\sim \mathrm{Ga}(0.001, 0.001), && && \text{(2.28e)} \end{aligned}$$

where 'N' denotes the scalar Gaussian (Normal) distribution, with specified expectation and variance, and 'Ga' denotes the Gamma distribution with specified shape and rate. The hyperparameters are $\Psi = (\mu, \tau^2)$ and the additional dispersion parameter is σ^2. As can be inferred, the convention with the extensive form of a hierarchical model is that the joint distribution is constructed by taking the product over the rows, and within rows by taking the product over each index. The extensive form is used in programming languages such as BUGS, JAGS, and STAN.

The last three lines in (2.28) represent something of a fudge. The three parameters μ, τ^2, and σ^2 are modelled independently (hence appearing on separate rows). They are all three given 'flat' priors, in order to make them maximally responsive to the observations. The standard flat priors would be μ uniform on $\mathbb{R}$, and $\log \tau^2$ and $\log \sigma^2$ each uniform on $\mathbb{R}$.[4] But these are improper distributions. The choices above represent proper distributions which approximate the flat priors. It is hard to recommend this approach, especially in the light of prior beliefs about X which would constrain the priors. But it is widely used, and there is little harm as long as the statistician checks carefully for the influence of the prior distribution on the posterior predictions.

Hierarchical modelling is a very rich approach to statistical modelling, and the illustration above is about as simple as they come—most hierarchical models are more interesting than (2.28), particularly in incorporating covariates, which typically enter at the top level. See Lunn et al. (2013) or Gelman et al. (2014) for more details. Hierarchical models are also used for X's with more complex structure than 'similar but not identical'; see Banerjee et al. (2004) or Cressie and Wikle (2011) for applications in spatial and spatio-temporal statistics.

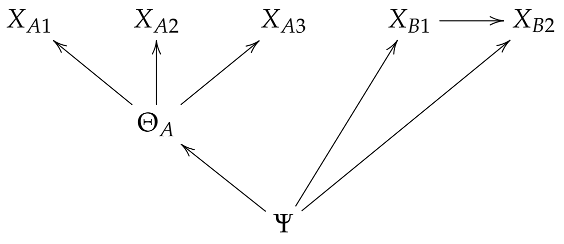
* * *

The illustration provides an opportunity to see again how the maxim of statistical modelling (Definition 2.1) works. Our intention is to construct an interesting statistical model for X so that we can predict (X_3, X_5) from the observations $(x_1^{\text{obs}}, x_2^{\text{obs}}, x_4^{\text{obs}})$. We introduced some additional random quantities, namely (Ψ, Θ), and construct a statistical model for (Ψ, Θ, X) which exploits lots

[4] Technically, this is equivalent to $(\mu, \log \tau^2, \log \sigma^2)$ being uniform on $\mathbb{R}^3$.

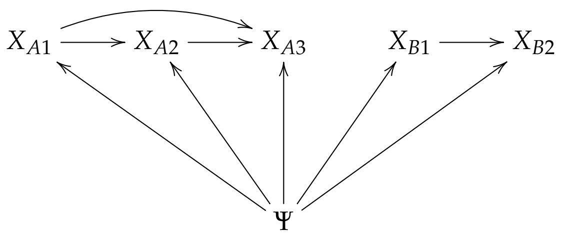
of conditional independence, based on our understanding of the group structure of X. Then, notionally at least, we recover the marginal distribution for X by marginalizing over (Ψ, Θ).

Theorem 2.4 needs to be adapted if we want to marginalize over, say, Θ_B, which is not at the bottom level of the DAG (i.e. it has parents). Following exactly the same logic as Theorem 2.4, each 'child' of Θ_B inherits the parents of Θ_B, and X_{B2} also gains an edge from X_{B1}, to give



(2.29)

If we also marginalize over Θ_A we get



(2.30)

And if we were now to marginalize over Ψ we would have a full set of 10 edges. The maxim of statistical modelling is simply that introducing random variables and marginalizing over them is a better way to construct a PMF for X than trying to write down such a PMF directly.

2.7 *Full conditionals*

In this section I will return to DAGs constructed for the PMF f_X, without distinguishing between random quantities and random variables; i.e. each X_i might be either a random quantity or a random variable.

It is a basic property of a DAG that it is constructed with respect to a specific ordering of all of the random quantities. We cannot simply glance at the DAG and answer questions such as "Is X_i conditionally independent of X_j given X_C?", for some arbitrary set C not containing i or j. Of course we could get lucky. Suppose that $i < j$, without loss of generality (by symmetry of conditional independence). If it so happened that $\text{pa}_j = C$ and $i \notin C$ then we would be able to answer "Yes"; otherwise, we do not know—at least, not without further thought.

To explore this issue, consider the challenge of finding the *full conditional* of X_i, which is defined as

$$p(x_i \mid x_{-i}) \tag{2.31}$$

where X_{-i} is every random quantity bar X_i. Suppose we found that the full conditional only depended on a subset of x_{-i}, say the index set 'ne$_i$'. In that case we would have shown that

$$X_i \perp\!\!\!\perp X_{\overline{\text{ne}}_i} \mid X_{\text{ne}_i} \tag{2.32}$$

where $\overline{\text{ne}}_i$ is the complement of $\{i\} \cup \text{ne}_i$ in $\{1, \dots, m\}$. Here 'ne$_i$' denotes the 'neighbours' of X_i, as explained below.

If we found the neighbours for each X_i we could construct a new graph on X: a graph in which there was an edge from X_i to X_j exactly when $i \in \text{ne}_j$. This is an undirected graph because

$$i \in \text{ne}_j \iff j \in \text{ne}_i \tag{2.33}$$

(proved below, Theorem 2.6). Call this the *conditional independence graph (CIG)* of f_X. The CIG of f_X has a remarkable property, given by the Hammersley-Clifford theorem.[5]

Theorem 2.5 (Hammersley-Clifford theorem)**.** *If*

$$\operatorname{supp} X = \prod_{i=1}^{m} \operatorname{supp} X_i,$$

then $X_A \perp\!\!\!\perp X_B \mid X_C$ if and only if every path on the CIG from X_A to X_B passes through X_C.[6]

In these notes I will not prove the Hammersley-Clifford theorem, but I will show how to turn a DAG into a CIG, so that every conditional independence property can be read off (subject to the positivity condition).

Theorem 2.6 (Moralization theorem)**.** *The following two steps transform the DAG of f_X into the CIG:*

1. *Insert an edge between every pair of vertices which share a child.*
2. *Replace all directed edges with undirected edges (i.e. remove arrows).*

Proof. I will write 'pa$_j$' in place of 'x_{pa_j}', for clarity. The full conditional of X_i is

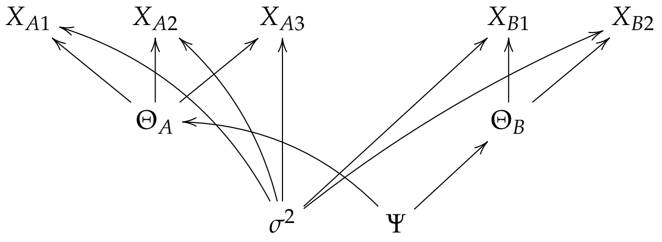
$$\begin{aligned} \mathrm{p}(x_i \mid x_{-i}) &= \frac{\mathrm{p}(x)}{\sum_{x_i} \mathrm{p}(x)} \\ &= \frac{\prod_{j=1}^{m} \mathrm{p}(x_j \mid \text{pa}_j)}{\sum_{x_i} \prod_{j=1}^{m} \mathrm{p}(x_j \mid \text{pa}_j)} \\ &\propto p(x_i \mid \text{pa}_i) \cdot \prod_{j \in A_i} p(x_j \mid \text{pa}_j) \end{aligned} \tag{2.34}$$

where $\text{pa}_1 = \emptyset$ and $A_i := \{j : i \in \text{pa}_j\}$. The last step follows because in the denominator, terms without i in pa$_j$ come through the sum and then cancel with the same terms in the numerator. This leaves only $\mathrm{p}(x_i \mid \text{pa}_i)$ and terms with $i \in \text{pa}_j$. This comprises all parents of X_i, all children of X_i, and all vertices which share a child with X_i, which proves (2.33), and justifies the two steps in the theorem. □

[5] Which some authors refer to as the Hammersley-Clifford-Besag theorem, following Besag (1974). This is not a trivial theorem; see Besag (1974) for one proof, and Lauritzen (1996, ch. 3) for another. Besag (1974) is one of the great papers in modern statistics.

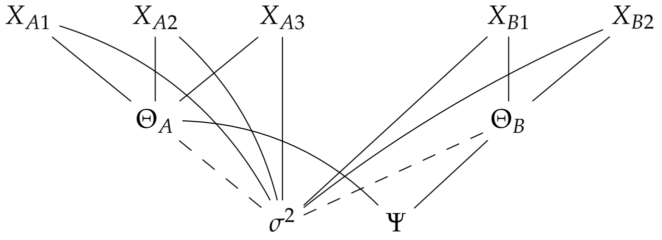
[6] The 'support' of X is $\operatorname{supp} X := \{x \in \mathcal{X} : f_X(x) > 0\}$. The condition that the support of X is equal to the product of the marginal supports is termed the *positivity condition*; it cannot be dropped.

The CIG of (2.25) is the same as the DAG, but without the arrows. This is because (2.25) is a tree, which is a directed graph in which every vertex has exactly one parent. But if we add in σ^2, the DAG becomes



(2.35)

which is a bit messy. Careful inspection reveals two new edges in the CIG, shown below as dashed:



(2.36)

Applying the Hammersley-Clifford theorem (Theorem 2.5) we can read off, for example,

$$X_{A1} \perp\!\!\!\perp \text{all other } X\text{'s} \mid \Theta_A, \sigma^2$$
$$X_A \perp\!\!\!\perp X_B \mid \Theta_A, \sigma^2$$

subject to the positivity condition, which is satisfied in (2.28) because the support of each random quantity/variable does not depend on its parameters.