score-based structure learning in Section 4.3 and constraint-based structure learning in Section 4.4. In Section 4.5 we apply structure learning to inferring causal influences from data. Section 4.6 presents learning packages that implement the methods discussed in the first three sections. Finally, in Section 4.7 we show examples of learning Bayesian networks and of causal learning.

4.1 Parameter Learning

We can only learn parameters from data when the probabilities are relative frequencies, which were discussed in Chapter 2, Section 2.3.1. So this discussion pertains only to such probabilities. Although the method is based on rigorous mathematical results obtained by modeling an individual's subjective belief concerning a relative frequency, the method itself is quite simple. Here, we merely present the method. See [Neapolitan, 2004] for the mathematical development. First, we discuss learning a single parameter, and then we show how to learn all the parameters in a Bayesian network.

4.1.1 Learning a Single Parameter

After presenting a method for learning the probability of a binomial variable, we extend the method to multinomial variables. Finally, we provide guidelines for articulating our prior beliefs concerning probabilities.

Binomial Variables

We illustrate learning with a sequence of examples.

Example 4.1 *Recall the discussion concerning a thumbtack at the beginning of Chapter 2, Section 2.3.1. We noted that a thumbtack could land on its flat end, which we call "heads," or it could land with the edge of the flat end and the point touching the ground, which we call "tails." Because the thumbtack is not symmetrical, we have no reason to apply the Principle of Indifference and assign probabilities of* .5 *to both outcomes. So we need data to estimate the probability of heads. Suppose we toss the thumbtack* 100 *times, and it lands heads* 65 *of those times. Then we can estimate*

$$P(heads) \approx \frac{65}{100} = .65.$$

This estimate, obtained by dividing the number of heads by the number of trials, is called the **maximum likelihood estimate (MLE)** of the probability. In general, if there are s heads in n trials, the MLE of the probability is

$$P(heads) \approx \frac{s}{n}.$$

Using the MLE seems reasonable when we have no prior belief concerning the probability. However, it is not so reasonable when we do have prior belief. Consider the next example.

Example 4.2 *Suppose you take a coin from your pocket, toss it* 10 *times, and it lands heads all of those times. Then using the MLE we estimate*

$$P(heads) \approx \frac{10}{10} = 1.$$

After the coin landed heads 10 times, we would not bet as if we were certain that the outcome of the 11th toss will be heads. So our belief concerning the $P(heads)$ is not the MLE value of 1. Assuming we believe the coins in our pockets are fair, should we instead maintain $P(heads) = .5$ after all 10 tosses landed heads? This may seem reasonable for 10 tosses, but it does not seem so reasonable if 1000 straight tosses landed heads. At some point we would start suspecting the coin was weighted to land heads. We need a method that incorporates one's prior belief with the data. The standard way to do this is for the probability assessor to ascertain integers a and b such that the assessor's experience is equivalent to having seen the first outcome (heads in this case) occur a times and the second outcome occur b times in $m = a + b$ trials. Then the assessor's prior probabilities are

$$P(heads) = \frac{a}{m} \qquad P(tails) = \frac{b}{m}.$$

After observing s heads and t tails in $n = s + t$ trials, the assessor's posterior probabilities are

$$P(heads|s,t) = \frac{a+s}{m+n} \qquad P(tails|s,t) = \frac{a+t}{m+n}. \tag{4.1}$$

This probability is called the **maximum a posterior probability (MAP)**. Note that we have used the symbol "=" rather than "≈," and we have written the probability as a conditional probability rather than as an estimate. The reason is that this is a Bayesian technique, and Bayesians say that the value is their probability (belief) based on the data rather than saying it is an estimate of a probability (relative frequency). It is the assumption of exchangeability that enables us to update our beliefs using Equality 4.1. The assumption of **exchangeability**, which was first developed by de Finetti in 1937, is that an individual assigns the same probability to all sequences of the same length containing the same number of each outcome. For example, the individual assigns the same probability to these two sequences of heads (H) and tails (T):

$$H, T, H, T, H, T, H, T, T, T \qquad \text{and} \qquad H, T, T, T, T, H, H, T, H, T.$$

Furthermore, the individual assigns the same probability to any other sequence of 10 tosses that has 4 heads and 6 tails. Neapolitan [2004] discusses exchangeability in detail and derives Equality 4.1. Here, we accept that equality on intuitive grounds.

Next, we show more examples. In these examples we only compute the probability of the first outcome because the probability of the second outcome is uniquely determined by it.

Example 4.3 *Suppose you are going to repeatedly toss a coin from your pocket. Since you would feel it highly probable that the relative frequency is around* .5, *you might feel your prior experience is equivalent to having seen* 50 *heads in* 100 *tosses. Therefore, you could represent your belief with* $a = 50$ *and* $b = 50$. *Then* $m = 50 + 50 = 100$, *and your prior probability of heads is*

$$P(heads) = \frac{a}{m} = \frac{50}{100} = .5.$$

After seeing 48 *heads in* 100 *tosses, your posterior probability is*

$$P(heads|48, 52) = \frac{a+s}{m+n} = \frac{50+48}{100+100} = .49.$$

The notation 48, 52 *on the right of the conditioning bar in* $P(heads|48, 52)$ *represents the event that* 48 *heads and* 52 *tails have occurred.*

Example 4.4 *Suppose you are going to repeatedly toss a thumbtack. Based on its structure, you might feel it should land heads about half the time, but you are not nearly so confident as you were with the coin from your pocket. So you might feel your prior experience is equivalent to having seen* 3 *heads in* 6 *tosses. Then your prior probability of heads is*

$$P(heads) = \frac{a}{m} = \frac{3}{6} = .5.$$

After seeing 65 *heads in* 100 *tosses, your posterior probability is*

$$P(heads|65{,}35) = \frac{a+s}{m+n} = \frac{3+65}{6+100} = .64.$$

Example 4.5 *Suppose you are going to sample individuals in the United States and determine whether they brush their teeth. In this case, you might feel your prior experience is equivalent to having seen* 18 *individuals brush their teeth out of* 20 *sampled. Then your prior probability of brushing is*

$$P(brushes) = \frac{a}{m} = \frac{18}{20} = .9.$$

After sampling 100 *individuals and learning* 80 *brush their teeth, your posterior probability is*

$$P(brushes|80, 20) = \frac{a+s}{m+n} = \frac{18+80}{20+100} = .82.$$

You may feel that if we have complete prior ignorance as to the probability we should take $a = b = 0$. However, consider the next example.

Example 4.6 *Suppose we are going to sample dogs and determine whether or not they eat the potato chips which we offer them. Since we have no idea whether a particular dog would eat potato chips, we assign* $a = b = 0$, *which* *means $m = 0 + 0 = 0$. Since we cannot divide a by m, we have no prior probability. Suppose next that we sample one dog, and that dog eats the potato chips. Our probability of the next dog eating potato chips is now*

$$P(eats|1,0) = \frac{a+s}{m+n} = \frac{0+1}{0+1} = 1.$$

This belief is not very reasonable as we are now certain that all dogs eat potato chips. Owing to difficulties such as this and more rigorous mathematical results, prior ignorance to a probability is usually modeled by taking $a = b = 1$, which means $m = 1 + 1 = 2$. If we use these values instead, our posterior probability when the first dog sampled was found to eat potato chips is given by

$$P(eats|1,0) = \frac{a+s}{m+n} = \frac{1+1}{2+1} = \frac{2}{3}.$$

Sometimes we want fractional values for a and b. Consider this example.

Example 4.7 *This example is taken from [Berry, 1996]. Glass panels in high-rise buildings sometimes break and fall to the ground. A particular case involved 39 broken panels. In their quest for determining why the panels broke, the owners wanted to analyze the broken panels, but they could only recover three of them. These three were found to contain Nickel Sulfide (NiS), a manufacturing flaw which can cause panels to break. In order to determine whether they should hold the manufacturer responsible, the owners then wanted to determine how probable it was that all 39 panels contained NiS. So they contacted a glass expert.*

The glass expert testified that among glass panels that break, only 5% contain NiS. However, NiS tends to be pervasive in production lots. So given that the first panel sampled, from a particular production lot of broken panels, contains NiS, the expert felt the probability was .95 that the second panel sampled also contains NiS. It was known that all 39 panels came from the same production lot. So, if we model the expert's prior belief using values of a, b, and $m = a + b$ as discussed above, we must have that the prior probability is given by

$$P(NiS) = \frac{a}{m} = .05.$$

Furthermore, the expert's posterior probability after finding the first panel contains NiS must be given by

$$P(NiS|1,0) = \frac{a+1}{m+1} = .95.$$

Solving these last two equations for a and m yields

$$a = \frac{1}{360} \qquad m = \frac{20}{360}.$$

This is an alternative technique for assessing a and b. Namely, we assess the probability for the first trial. Then we assess the conditional probability for the *second trial given the first one is a "success." Once we have values of a and b, we can determine how likely it is that any one of the other 36 panels (the next one sampled) contains NiS after the first three sampled were found to contain it. We have that this probability is given by*

$$P(NiS|3,0) = \frac{a+s}{m+n} = \frac{1/360+3}{20/360+3} = .983.$$

Notice how the expert's probability of NiS quickly changed from being very small to be being very large. This is because the values of a and m are so small. We are really most interested in whether all 36 remaining panels contain NiS. It is left as an exercise to show that this probability is given by

$$\prod_{i=0}^{35} \frac{1/360+3+i}{20/360+3+i} = .866.$$

Multinomial Variables

The method just discussed readily extends to multinomial variables. Suppose k is the number of values the variable can assume, and $x_1, x_2, \ldots, x_k$ are the k outcomes. We then ascertain numbers $a_1, a_2, \ldots, a_k$ such that our experience is equivalent to having seen the first outcome occur a_1 times, the second outcome occur a_2 times, ..., and the last outcome occur a_k times. Our prior probabilities before any trials are then

$$P(x_1) = \frac{a_1}{m} \qquad P(x_2) = \frac{a_2}{m} \qquad \cdots \qquad P(x_k) = \frac{a_k}{m},$$

where $m = a_1 + a_2 + \cdots + a_k$. After seeing x_1 occur s_1 times, x_2 occur s_2 times, ..., and x_n occur s_n times in $n = s_1 + s_2 + \cdots + s_k$ trials, our posterior probabilities are as follows:

$$\begin{aligned}
P(x_1|s_1, s_2, \ldots, s_k) &= \frac{a_1 + s_1}{m+n} \\
P(x_2|s_1, s_2, \ldots, s_k) &= \frac{a_2 + s_2}{m+n} \\
&\vdots \\
P(x_k|s_1, s_2, \ldots, s_k) &= \frac{a_k + s_k}{m+n}.
\end{aligned}$$

Example 4.8 *Suppose we have an asymmetrical-sided six-sided die, and we have little idea of the probability of each side coming up. However, it seems that all sides are equally likely. So we assign*

$$a_1 = a_2 = \cdots = a_6 = 3.$$

Then our prior probabilities are as follows:

$$P(1) = P(2) = \cdots = P(6) = \frac{a_i}{n} = \frac{3}{18} = .16667.$$

Suppose next we throw the die 100 times with the following result:

Outcome	Number of Occurrences
1	10
2	15
3	5
4	30
5	13
6	27

We then have

$$\begin{aligned}
P(1|10,15,5,30,13,27) &= \frac{a_1+s_1}{m+n} = \frac{3+10}{18+100} = .110 \\
P(2|10,15,5,30,13,27) &= \frac{a_2+s_2}{m+n} = \frac{3+15}{18+100} = .153 \\
P(3|10,15,5,30,13,27) &= \frac{a_3+s_3}{m+n} = \frac{3+5}{18+100} = .067 \\
P(4|10,15,5,30,13,27) &= \frac{a_4+s_4}{m+n} = \frac{3+30}{18+100} = .280 \\
P(5|10,15,5,30,13,27) &= \frac{a_5+s_5}{m+n} = \frac{3+13}{18+100} = .136 \\
P(6|10,15,5,30,13,27) &= \frac{a_6+s_6}{m+n} = \frac{3+27}{18+100} = .254.
\end{aligned}$$

Guidelines for Articulating Prior Belief

Next, we give some guidelines for choosing the values that represent our prior beliefs.

Binary Variables The guidelines for binary variables are as follows.

1. $a = b = 1$: We use these values when we feel we have no knowledge at all concerning the value of the probability. We might also use these values to try to achieve objectivity in the sense that we impose none of our beliefs concerning the probability on the learning algorithm. We only impose the fact that we know, at most, two things can happen. An example might be when we are learning the probability of lung cancer given smoking from data, and we want to communicate our result to the scientific community. The scientific community would not be interested in our prior belief; rather it would be interested only in what the data had to say. Essentially, when we use these values, the posterior probability represents the belief of an individual who has no prior belief concerning the probability.

2. $a, b > 1$: These values mean we feel it is likely that the probability of the first outcome is a/m. The larger the values of a and b are, the more we believe this. We would use such values when we want to impose our beliefs concerning the relative frequency on the learning algorithm. For

example, if we were going to toss a coin taken from the pocket, we might take $a = b = 50$.

3. $a, b < 1$: These values mean we feel it is likely that the probability of one of the outcomes is high, although we are not committed to which one it is. If we take $a = b \approx 0$ (almost 0), then we are almost certain the probability of one of the outcomes is very close to 1. We would also use values like these when we want to impose our beliefs concerning the probability on the learning algorithm. Example 4.7 shows a case in which we would choose values less than 1. Notice that such prior beliefs are quickly overwhelmed by data. For example, if $a = b = .1$, and we saw the first outcome x_1 occur in a single trial, we have

$$P(x_1|1,0) = \frac{.1+1}{.2+1} = .917. \tag{4.2}$$

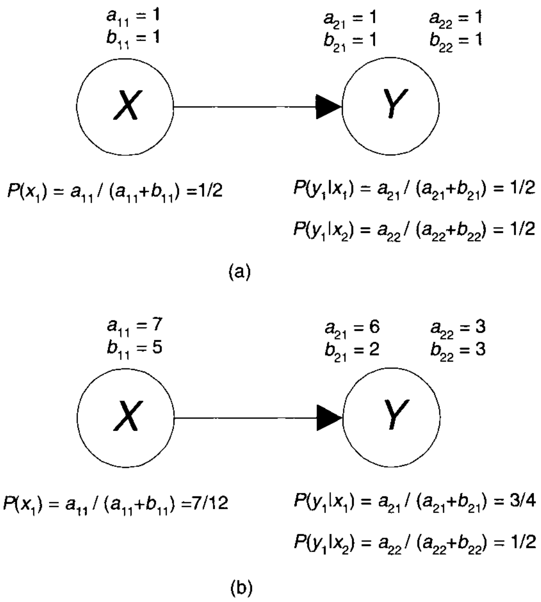
Intuitively, we thought *a priori* that the probability of one of the outcomes was high. The fact that it took the value x_1 once makes us believe it is probably that outcome.

Multinomial Variables The guidelines are essentially the same as those for binomial variables, but we restate them for the sake of clarity.

1. $a_1 = a_2 = \cdots = a_k = 1$: We use these values when we feel we have no knowledge at all concerning the probabilities. We might also use these values to try to achieve objectivity in the sense that we impose none of our beliefs concerning the probability on the learning algorithm. We only impose the fact that we know, at most, k things can happen. An example might be learning the probability of low, medium, and high blood pressure from data, which we want to communicate to the scientific community.

2. $a_1 = a_2 = \cdots = a_k > 1$: These values mean we feel it more likely that the probability of the kth value is around a_k/m. The larger the values of a_k are, the more we believe this. We would use such values when we want to impose our beliefs concerning the probability on the learning algorithm. For example, if we were going to toss an ordinary die, we might take $a_1 = a_2 = \cdots = a_6 = 50$.

3. $a_1 = a_2 = \cdots = a_k < 1$: These values mean we feel that it is likely that only a few outcomes are probable. We would use such values when we want to impose our beliefs concerning the probabilities on the learning algorithm. For example, suppose we know there are 1,000,000 different species in the world, and we are about to land on an uncharted island. We might feel it probable that not very many of the species are present. So if we considered the probabilities with which we encountered different species, we would not consider probabilities, which resulted in a lot of different species, likely. Therefore, we might take $a_i = 1/1{,}000{,}000$ for all i.


$a_{11} = 1$
$b_{11} = 1$
$a_{21} = 1$
$b_{21} = 1$
$a_{22} = 1$
$b_{22} = 1$
X
Y
$P(x_1) = a_{11} / (a_{11}+b_{11}) = 1/2$
$P(y_1|x_1) = a_{21} / (a_{21}+b_{21}) = 1/2$
$P(y_1|x_2) = a_{22} / (a_{22}+b_{22}) = 1/2$
$a_{11} = 7$
$b_{11} = 5$
$a_{21} = 6$
$b_{21} = 2$
$a_{22} = 3$
$b_{22} = 3$
X
Y
$P(x_1) = a_{11} / (a_{11}+b_{11}) = 7/12$
$P(y_1|x_1) = a_{21} / (a_{21}+b_{21}) = 3/4$
$P(y_1|x_2) = a_{22} / (a_{22}+b_{22}) = 1/2$

(a)

(b)

Figure 4.1: A Bayesian network initialized for parameter learning appears in (a), while the updated network based on the data in Figure 4.2 appears in (b).

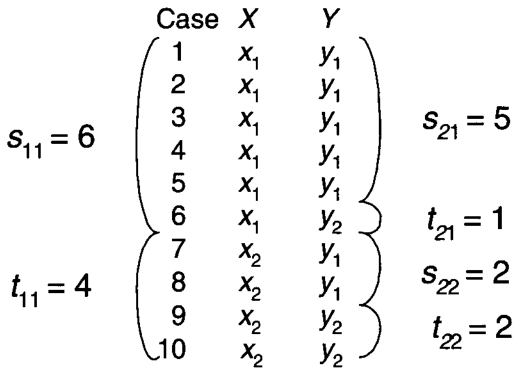
4.1.2 Learning All Parameters in a Bayesian Network

The method for learning all parameters in a Bayesian network follows readily from the method for learning a single parameter. We illustrate the method with binomial variables. It extends readily to the case of multinomial variables (see [Neapolitan, 2004]). After showing the method, we discuss equivalent sample sizes.

Procedure for Learning Parameters

Consider the two-node Bayesian network in Figure 4.1 (a). It has been initialized for parameter learning. For each probability in the network there is a pair (a_{ij}, b_{ij}). The i indexes the variable, while the j indexes the value of the parent(s) of the variable. For example, the pair (a_{11}, b_{11}) is for the 1st variable (X), and the 1st value of its parent (in this case there is a default of one parent value since X has no parent). The pair (a_{21}, b_{21}) is for the 2nd variable (Y), and the 1st value of its parent, namely x_1. The pair (a_{22}, b_{22}) is for the 2nd variable (Y), and the 2nd value of its parent, namely x_2. We have attempted to represent prior ignorance as to the value of all probabilities by taking $a_{ij} = b_{ij} = 1$. We compute the prior probabilities using these pairs just as we did when we were considering a single parameter. We have

$$P(x_1) = \frac{a_{11}}{a_{11} + b_{11}} = \frac{1}{1+1} = \frac{1}{2}$$



Case	X	Y
1	x_1	y_1
2	x_1	y_1
3	x_1	y_1
4	x_1	y_1
5	x_1	y_1
6	x_1	y_2
7	x_2	y_1
8	x_2	y_1
9	x_2	y_2
10	x_2	y_2

$s_{11} = 6$
$t_{11} = 4$
$s_{21} = 5$
$t_{21} = 1$
$s_{22} = 2$
$t_{22} = 2$

Figure 4.2: Data on 10 cases.

$$P(y_1|x_1) = \frac{a_{21}}{a_{21} + b_{21}} = \frac{1}{1+1} = \frac{1}{2}$$

$$P(y_1|x_2) = \frac{a_{22}}{a_{22} + b_{22}} = \frac{1}{1+1} = \frac{1}{2}.$$

When we obtain data, we use an (s_{ij}, t_{ij}) pair to represent the counts for the ith variable when the variable's parents have their jth value. Suppose we obtain the data in Figure 4.2. The values of the (s_{ij}, t_{ij}) pairs are shown in that figure. We have that $s_{11} = 6$ because x_1 occurs 6 times, and $t_{11} = 4$ because x_2 occurs 4 times. Of the 6 times that x_1 occurs, y_1 occurs 5 times and y_2 occurs 1 time. So $s_{21} = 5$ and $t_{21} = 12$. Of the 4 times that x_2 occurs, y_1 occurs 2 times and y_2 occurs 2 times. So $s_{22} = 2$ and $t_{22} = 2$. To determine the posterior probability distribution based on the data we update each conditional probability with the counts relative to that conditional probability. Since we want an updated Bayesian network, we recompute the values of the (a_{ij}, b_{ij}) pairs. We therefore have

$$\begin{aligned}
a_{11} &= a_{11} + s_{11} = 1 + 6 = 7 \\
b_{11} &= b_{11} + t_{11} = 1 + 4 = 5 \\
\\
a_{21} &= a_{21} + s_{21} = 1 + 5 = 6 \\
b_{21} &= b_{21} + t_{21} = 1 + 1 = 2 \\
\\
a_{22} &= a_{22} + s_{22} = 1 + 2 = 3 \\
b_{22} &= b_{22} + t_{22} = 1 + 2 = 3.
\end{aligned}$$

We then compute the new values of the parameters:

$$P(x_1) = \frac{a_{11}}{a_{11} + b_{11}} = \frac{7}{7+5} = \frac{7}{12}$$

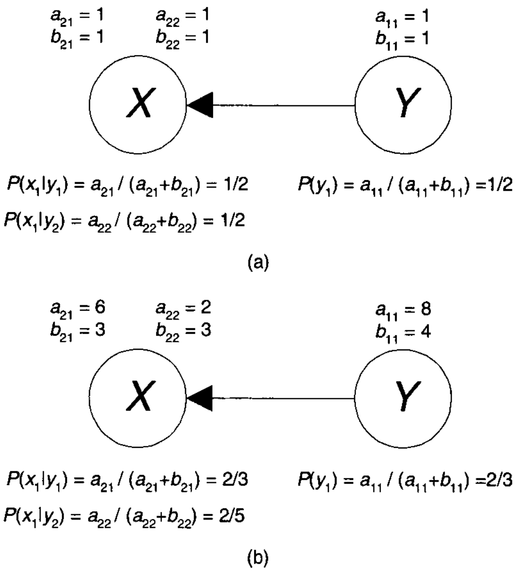

a_{21} = 1
b_{21} = 1
a_{22} = 1
b_{22} = 1
a_{11} = 1
b_{11} = 1
X
Y
$P(x_1|y_1) = a_{21} / (a_{21}+b_{21}) = 1/2$
$P(x_1|y_2) = a_{22} / (a_{22}+b_{22}) = 1/2$
$P(y_1) = a_{11} / (a_{11}+b_{11}) = 1/2$
(a)
a_{21} = 6
b_{21} = 3
a_{22} = 2
b_{22} = 3
a_{11} = 8
b_{11} = 4
X
Y
$P(x_1|y_1) = a_{21} / (a_{21}+b_{21}) = 2/3$
$P(x_1|y_2) = a_{22} / (a_{22}+b_{22}) = 2/5$
$P(y_1) = a_{11} / (a_{11}+b_{11}) = 2/3$
(b)

Figure 4.3: A Bayesian network initialized for parameter learning appears in (a), while the updated network based on the data in Figure 4.2 appears in (b).

$$P(y_1|x_1) = \frac{a_{21}}{a_{21}+b_{21}} = \frac{6}{6+2} = \frac{3}{4}$$

$$P(y_1|x_2) = \frac{a_{22}}{a_{22}+b_{22}} = \frac{3}{3+3} = \frac{1}{2}.$$

The updated network is shown in Figure 4.1 (b).

Equivalent Sample Size

There is a problem with the way we represented prior ignorance in the preceding subsection. Although it seems natural to set $a_{ij} = b_{ij} = 1$ to represent prior ignorance of all the conditional probabilities, such assignments are not consistent with the metaphor we used for articulating these values. Recall that we said the probability assessor is to choose values of a and b such that the assessor's experience is equivalent to having seen the first outcome occur a times in $a + b$ trials. Therefore, if we set $a_{11} = b_{11} = 1$, the assessor's experience is equivalent to having seen x_1 occur 1 time in 2 trials. However, if we set $a_{21} = b_{21} = 1$, the assessor's experience is equivalent to having seen y_1 occur 1 time out of the 2 times x_1 occurred. This is not consistent. First, we are saying x_1 occurred once; then we are saying it occurred twice. Aside from this inconsistency, we obtain odd results if we use these priors. Consider the Bayesian network for parameter learning in Figure 4.3 (a). If we update that network with the data in Figure 4.2, we obtain the network in Figure 4.3 (b). The DAG in Figure 4.3 (a) is equivalent to the one in Figure 4.1 (a) in a certain sense. That is, we say two DAGs are **Markov equivalent** if they entail the same conditional independencies. The DAGs in Figures 4.1 (a) and 4.3 (a) are Markov equivalent because both DAGs entail no conditional independencies. It seems that if we represent the same prior beliefs with equivalent DAGs, then the posterior distributions based on data should be the same. In this case we have attempted to represent prior ignorance as to all probabilities with the networks in Figure 4.1 (a) and Figure 4.3 (a). So the posterior distributions based on the data in Figure 4.2 should be the same. However, from the Bayesian network in Figure 4.1 (b) we have

$$P(x_1) = \frac{7}{12} = .583,$$

while from the Bayesian network in Figure 4.3 (b) we have

$$\begin{aligned} P(x_1) &= P(x_1|y_1)P(y_1) + P(x_1|y_2)P(y_2) \\ &= \frac{2}{3} \times \frac{2}{3} + \frac{2}{5} \times \frac{1}{3} = .578. \end{aligned}$$

We see that we obtain different posterior probabilities. Such results are not only odd, but unacceptable since we have attempted to model the same prior belief with the Bayesian networks in Figures 4.1 (a) and 4.3 (a), but end up with different posterior beliefs.

We can eliminate this difficulty by using a prior equivalent sample size. That is, we specify values of a_{ij} and b_{ij} that could actually occur in a prior sample that exhibit the conditional independencies entailed by the DAG. For example, given the network $X \rightarrow Y$, if we specify that $a_{21} = b_{21} = 1$, this means our prior sample must have x_1 occurring 2 times. So we need to specify $a_{11} = 2$. Similarly, if we specify that $a_{22} = b_{22} = 1$, this means our prior sample must have x_2 occurring 2 times. So we need to specify $b_{11} = 2$. Note that we are not saying we actually have a prior sample. We are saying the probability assessor's beliefs are represented by a prior sample. Figure 4.4 shows prior Bayesian networks using equivalent sample sizes. Notice that the values of a_{ij} and b_{ij} in these networks represent the following prior sample:

Case	X	Y
1	x_1	y_1
2	x_1	y_2
3	x_2	y_1
4	x_2	y_2

It is left as an exercise to show that if we update both the Bayesian networks in Figure 4.4 using the data in Figure 4.2, we obtain the same posterior probability distribution. This result is true, in general. We state it as a theorem, but first give a formal definition of a prior equivalent sample size.

Definition 4.1 *Suppose we specify a Bayesian network for the purpose of learning parameters in the case of binomial variables. If there is a number*

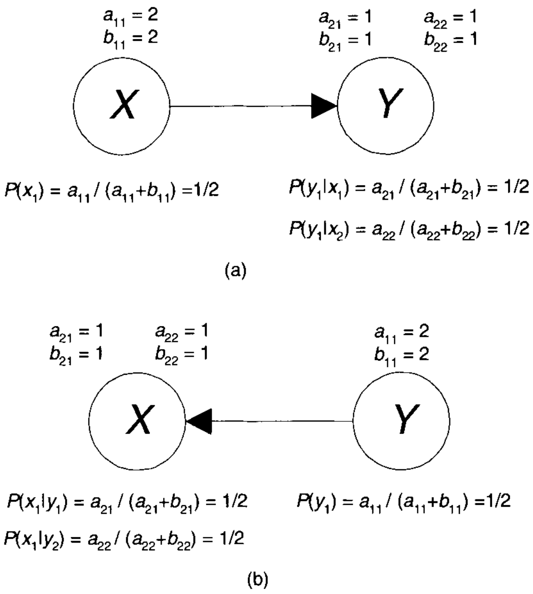

$a_{11} = 2$
$b_{11} = 2$
$a_{21} = 1$
$b_{21} = 1$
$a_{22} = 1$
$b_{22} = 1$
X
Y
$P(x_1) = a_{11} / (a_{11}+b_{11}) = 1/2$
$P(y_1|x_1) = a_{21} / (a_{21}+b_{21}) = 1/2$
$P(y_1|x_2) = a_{22} / (a_{22}+b_{22}) = 1/2$
(a)
$a_{21} = 1$
$b_{21} = 1$
$a_{22} = 1$
$b_{22} = 1$
$a_{11} = 2$
$b_{11} = 2$
X
Y
$P(x_1|y_1) = a_{21} / (a_{21}+b_{21}) = 1/2$
$P(x_1|y_2) = a_{22} / (a_{22}+b_{22}) = 1/2$
$P(y_1) = a_{11} / (a_{11}+b_{11}) = 1/2$
(b)

Figure 4.4: Bayesian networks for parameter learning containing prior equivalent sample sizes.

N such that for all i and j

$$a_{ij} + b_{ij} = P(\mathsf{pa}_{ij}) \times N,$$

where pa_{ij} denotes the jth instantiation of the parents of the ith variable, then we say the network has prior ***equivalent sample size*** *N.*

This definition is a bit hard to grasp by itself. The following theorem, whose proof can be found in [Neapolitan, 2004], yields a way to represent uniform prior distributions, which is often what we want to do.

Theorem 4.1 *Suppose we specify a Bayesian network for the purpose of learning parameters in the case of binomial variables and assign for all i and j*

$$a_{ij} = b_{ij} = \frac{N}{2q_i}.$$

where N is a positive integer and q_i is the number of instantiations of the parents of the ith variable. Then the resultant Bayesian network has equivalent sample size N, and the joint probability distribution in the Bayesian network is uniform.

We can represent prior ignorance by applying the preceding theorem with $N = 2$. The next example illustrates the technique.

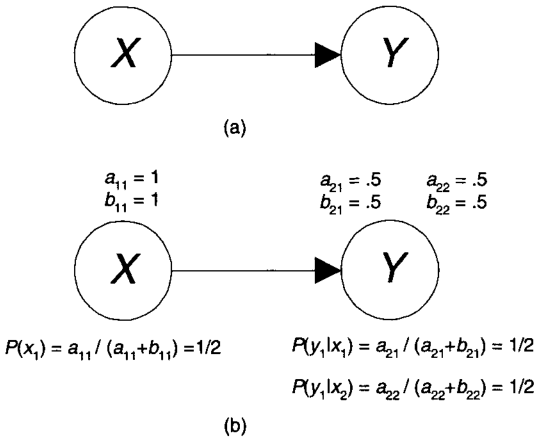

X
Y
(a)
$a_{11} = 1$
$b_{11} = 1$
$a_{21} = .5$
$b_{21} = .5$
$a_{22} = .5$
$b_{22} = .5$
X
Y
$P(x_1) = a_{11} / (a_{11}+b_{11}) = 1/2$
$P(y_1|x_1) = a_{21} / (a_{21}+b_{21}) = 1/2$
$P(y_1|x_2) = a_{22} / (a_{22}+b_{22}) = 1/2$
(b)

Figure 4.5: Given the DAG in (a) and that X and Y are binomial variables, the Bayesian network for parameter learning in (b) represents prior ignorance.

Example 4.9 *Suppose we start with the DAG in Figure 4.5 (a) and X and Y are binary variables. Set $N = 2$. Then using the method in Theorem 4.1 we have*

$$a_{11} = b_{11} = \frac{N}{2q_1} = \frac{2}{2 \times 1} = 1$$

$$a_{21} = b_{21} = a_{22} = b_{22} = \frac{N}{2q_2} = \frac{2}{2 \times 2} = .5.$$

We obtain the Bayesian network for parameter learning in Figure 4.5 (b).

Note that we obtained fractional values for a_{21}, b_{21}, a_{22}, and b_{22} in the preceding example, which may seem odd. However, the sum of these values is

$$a_{21} + b_{21} + a_{22} + b_{22} = .5 + .5 + .5 + .5 = 2.$$

So these fractional values are consistent with the metaphor that says we represent prior ignorance of the $P(Y)$ by assuming the assessor's experience is equivalent to having seen two trials (see Section 4.1.1). The following is an intuitive justification for why these values should be fractional. Recall from Section 4.1.1 that we said we use fractional values when we feel it is likely that the probability of one of the outcomes is high, although we are not committed to which one it is. The smaller the values are, the more likely we feel this is the case. Now the more parents a variable has the smaller are the values of a_{ij} and b_{ij} when we set $N = 2$. Intuitively, this seems reasonable because when a variable has many parents and we know the values of the parents, we know a lot about the state of the variable, and therefore, it is more likely the probability of one of the outcomes is high.

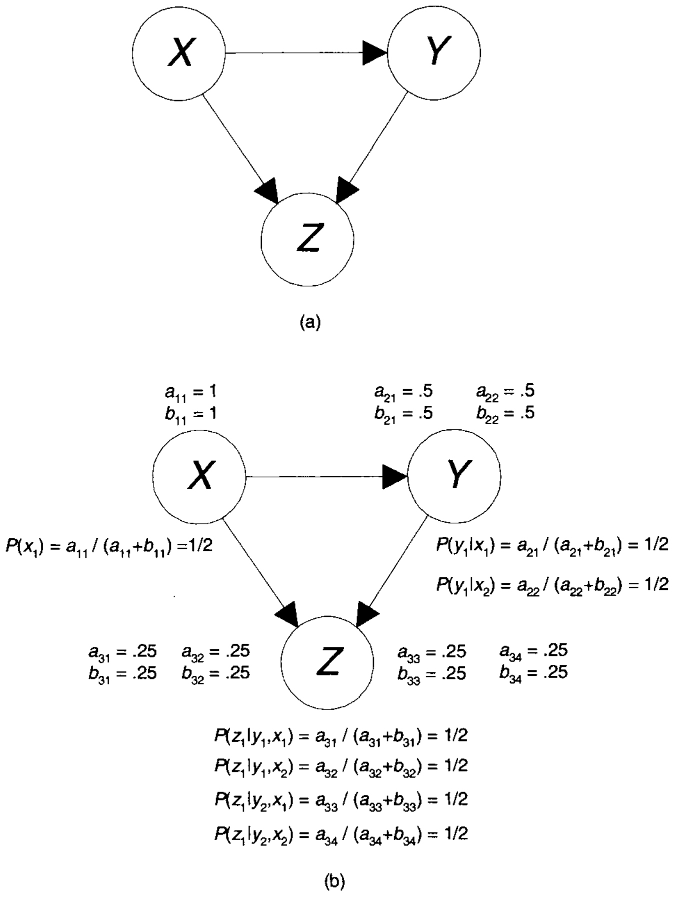

X
Y
Z
(a)
$a_{11} = 1$
$b_{11} = 1$
$a_{21} = .5$
$b_{21} = .5$
$a_{22} = .5$
$b_{22} = .5$
X
Y
$P(x_1) = a_{11} / (a_{11}+b_{11}) = 1/2$
$P(y_1|x_1) = a_{21} / (a_{21}+b_{21}) = 1/2$
$P(y_1|x_2) = a_{22} / (a_{22}+b_{22}) = 1/2$
$a_{31} = .25$
$b_{31} = .25$
$a_{32} = .25$
$b_{32} = .25$
Z
$a_{33} = .25$
$b_{33} = .25$
$a_{34} = .25$
$b_{34} = .25$
$P(z_1|y_1,x_1) = a_{31} / (a_{31}+b_{31}) = 1/2$
$P(z_1|y_1,x_2) = a_{32} / (a_{32}+b_{32}) = 1/2$
$P(z_1|y_2,x_1) = a_{33} / (a_{33}+b_{33}) = 1/2$
$P(z_1|y_2,x_2) = a_{34} / (a_{34}+b_{34}) = 1/2$
(b)

Figure 4.6: Given the DAG in (a) and that X, Y, and Z are binomial variables, the Bayesian network for parameter learning in (b) represents prior ignorance.

Example 4.10 *Suppose we start with the DAG in Figure 4.6 (a), and X and Y are binary variables. If we set $N = 2$ and use the method in Theorem 4.1, then*

$$a_{11} = b_{11} = \frac{N}{2q_1} = \frac{2}{2 \times 1} = 1$$

$$a_{21} = b_{21} = a_{22} = b_{22} = \frac{N}{2q_2} = \frac{2}{2 \times 2} = .5$$

$$a_{31} = b_{31} = a_{32} = b_{32} = a_{33} = b_{33} = a_{34} = b_{34} = \frac{N}{2q_3} = \frac{2}{2 \times 4} = .25.$$

We obtain the Bayesian network for parameter learning in Figure 4.6 (b).