

1

Introduction to Bayesian networks

Olivier Pourret

Electricité de France, 20 Place de la Défense, 92050, Paris la Défense, France

1.1 Models

1.1.1 Definition

The primary objective of this book is to discuss the power and limits of Bayesian networks for the purpose of constructing real-world models. The idea of the authors is not to extensively and formally expound on the formalism of mathematical models, and then explain that these models have been – or may be – applied in various fields; the point of view is, conversely, to explain why and how some recent, real-world problems have been modeled using Bayesian networks, and to analyse what worked and what did not.

Real-world problems – thus the starting point of this chapter – are often described as *complex*. This term is however seldom defined. It probably makes more sense to say that human cognitive abilities, memory, and reason are limited and that reality is therefore difficult to understand and manage. Furthermore, in addition to the biological limitations of human capabilities, a variety of factors, either cultural (education, ideology), psychological (emotions, instincts), and even physical (fatigue, stress) tend to distort our judgement of a situation.

One way of trying to better handle reality – in spite of these limitations and biases – is to use representations of reality called *models*. Let us introduce a basic example.

Example 1. Consider an everyday life object, such as a DVD recorder. The life cycle of the device includes its phases of design, manufacture, marketing, sale, use, possibly break down/repair, and disposal. The owner of a DVD recorder is involved in a temporally delimited part of its life cycle (i.e., when the object is in his/her living-room) and has a specific need: being able to use the device. The user's instruction manual of a DVD recorder is a description of the device, written in natural language, which exclusively aims at explaining how the device is to be operated, and is expressly dedicated to the owner: the manual does not include any internal description of the device.

In this example, the user's instruction manual is a *model* of the DVD recorder.

The 20 application chapters of this book provide numerous examples of models: models of organizations (Japanese electrical companies), of facts (criminal cases), of individuals (students in a robotics course, patients suffering from liver disorders), of devices (a sprinkler system), of places (potentially 'mineralized' geographic areas in India), of documents (texts of the parliament of Andalusia), of commodities (Chilean wines), or of phenomena (crime in the city of Bangkok, terrorism threats against US military assets). These parts of reality are material or immaterial: we will use the word 'objects' to refer to them.

These objects, which are delimited in time and space, have only one common point: at some stage of their life cycle (either before they actually occurred in reality, or in 'real-time' when they occurred, or after their occurrence) they have been modeled, and Bayesian networks have been employed to construct the model.

Example 1 suggests that the purpose of a model is to satisfy a need of some person or organization having a particular interest in one or several aspects of the object, but not in a comprehensive understanding of its properties. Using the terminology of corporate finance, we will refer to these individuals or groups of individuals with the word *stakeholders*. Examples of stakeholders include users, owners, operators, investors, authorities, managers, clients, suppliers, competitors. Depending on the role of the stakeholder, the need can be to:

- document, evaluate, operate, maintain the object;
- explain, simulate, predict, or diagnose its behavior;
- or – more generally – make decisions or prepare action regarding the object.

The very first benefit of the model is often to help the stakeholder to explicitly state his need: once a problem is explicitly and clearly expressed, it is sometimes not far from being solved.

The construction of a model involves the intervention of at least one human *expert* (i.e., someone with a practical or theoretical experience of the object), and is sometimes also based upon direct, uninterpreted observations of the object. Figure 1.1 illustrates this process: in Example 1, the object is the DVD recorder; the stakeholder is the owner of the device, who needs to be able to perform the installation, connections, setup and operation of the DVD recorder; the model is the user's instruction manual, which is based on the knowledge of some expert

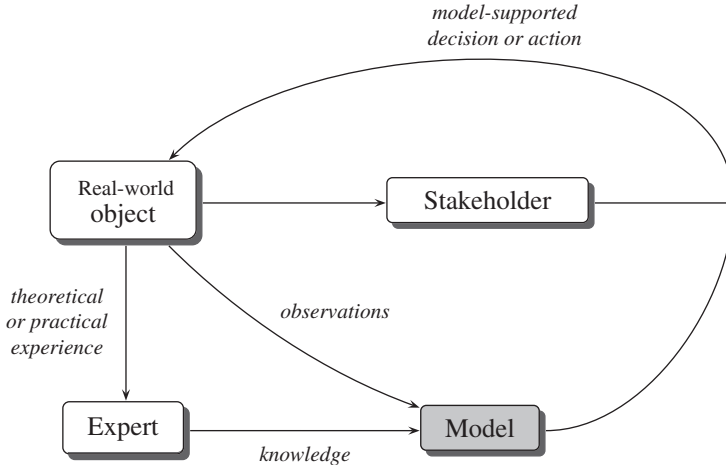


Figure 1.1 Construction and use of a model.

(i.e., the designer), and describes the object from the stakeholder’s point of view, in an understandable language; the model-supported action is the use of the device.

Based on the experience of the models discussed in this book, we may agree on the following definition – also fully consistent with Example 1.

Definition 2 (Model) *A model is a representation of an object, expressed in a specific language and in a usable form, and intended to satisfy one or several need(s) of some stakeholder(s) of the object.*

1.1.2 Use of a model: the inference process

Definition 2 specifies that models are written in a *usable* form. Let us analyse how models are used, i.e., explicitly state the *model-supported decision or action* arrow shown in Figure 1.1.

When the model includes an evaluation of the stakeholder’s situation, or a recommendation of decision or action, then the stakeholder makes his decision on the basis of the evaluation or follows the recommendation.

However, most models require – prior to their use – to be adapted to the specific situation of the stakeholder, by the integration of input data.

In Example 1, the input data are the type of the device, the information displayed by the device, and the actions already carried out by the user. The output information is the next action to be done.

Models are thus used to produce information (evaluations, appropriate decisions or actions) on the basis of some input information, considered as valid. This process is called *inference*.

Table 1.1 The inference process: given some input data, what can be inferred from the knowledge of the melting point of gold?

Input data	Inferred information
The ring is of solid gold. Temperature is 1000°C	The ring won't melt.
Temperature is 1000°C. The ring melts.	The ring is not of gold.
Temperature is 1100°C. The ring does not melt.	The ring is not of gold.
Temperature is 1100°C. The ring melts.	The ring is possibly of gold.
The ring is of solid gold. It does not melt.	The temperature is lower than T_m .

For example, if we assume that the statement

$$\text{The melting point of gold is } T_m = 1064.18^\circ\text{C} \tag{1.1}$$

is true, then it constitutes a model which can be used in numerous ways, depending on the available input data: Table 1.1 shows examples of information that can be inferred using the model, on the basis of some input data.

The use of real-world models is not always as straightforward as in the example of Table 1.1. For example, the model user may have some prior belief regarding whether the ring is of gold or not. Also, whether the rings melts or does not melt might be difficult to tell; finally, the temperature might not be known with a high level of precision. In such cases, the use of the model will not produce definitive ‘true’ statement, but just modify one’s assessment of the situation. For instance, if the ring is *believed* not to be of gold, the temperature *estimated* at 1100°C, and the ring *seems* not to melt, then the model output is that the ring is most unlikely of gold. If the uncertainties in the input data can be quantified with probabilities, then the use of the model increases the probability that the ring is not of gold. This is an example of *probabilistic* inference.

1.1.3 Construction

Definition 2 is extremely general: the way a model is constructed obviously depends on several factors, such as the nature of the object, the stakeholder’s need(s), the available knowledge and information, the time and resources devoted to the model elaboration, etc. Nevertheless, we may identify two invariants in the process of constructing a model.

1.1.3.1 Splitting the object into elements

One of the precepts of Descartes in his famous *Discourse on the Method* is ‘to divide each of the difficulties under examination into as many parts as possible and as might be necessary for its adequate solution’ [126].

Indeed, modeling an object implies splitting it into elements and identifying a number of aspects or attributes that characterise the elements.

Deriving a collection of attributes from one single object could at first glance appear as a poor strategy, but this really helps to simplify the problem of satisfying

the stakeholder's need: on one hand, each of the attributes is easier to analyze than the whole object; on the other hand, only the attributes which are relevant for the need of the stakeholder are taken into consideration.

1.1.3.2 Saying how it works: the modeling language

To allow inference, the model must include a description of how the elements interact and influence each other. As said in Definition 2, this involves the use of a specific language, which is either textual (natural language, formal rules), mathematical (equations, theorems), or graphical (plans, maps, diagrams).

The well-known consideration 'A good drawing is better than a long speech' also applies to models. Figures are more easily and quickly interpreted, understood and memorized than words. Models which are represented or at least illustrated in a graphical form tend to be more popular and commonly used. It is possible to admit that, throughout history, most successful or unsuccessful attempts of mankind to overcome the complexity of reality have involved, at some stage, a form a graphical representation. Human paleolithic cave paintings – although their interpretation in terms of hunting magic is not beyond dispute – may be considered as the first historical models, in the sense of Definition 2.

1.2 Probabilistic vs. deterministic models

1.2.1 Variables

During the modeling process, the exact circumstances in which the model is going to be used (especially, what input data the model will process) are, to a large extent, unknown. Also, some of the attributes remain unknown when the model is used: the attributes which are at some stage unknown are more conveniently described by *variables*.

In the rest of the chapter, we therefore consider an object which is characterized by a collection of numerical or symbolic variables, denoted X_1, X_2, \dots, X_n . To simplify the formalism, we suppose that the domain of each of the X_j variables, denoted \mathcal{E}_j , is discrete.

One may basically distinguish two kinds of variables. The first category is the variables whose values are specified (typically by the stakeholder) at some stage of the model use. Such variables typically describe:

- some aspects of the context: the values of such variables are defined prior to the use of the model and do not change afterwards (in Example 1: which version of the DVD recorder is being installed?);
- some aspects of the object which are directly 'controlled' by the stakeholder:
 - attributes of the object the stakeholder can observe (in Example 1: what is displayed on the control screen of the device?);

- decisions or actions he or she could carry out (in Example 1: what button should be pressed?).

The second category of variables are those which are not directly or not completely controlled – although possibly influenced – by the stakeholder’s will (in Example 1: Is the device well setup/out of order?).

At any stage of the practical use of a model, the variables under control have their value fixed, and do not explicitly behave as variables anymore. We may therefore suppose without any loss of generality that the model only comprises variables which are not under control.

1.2.2 Definitions

A deterministic model is a collection of statements, or rules regarding the X_i variables. A sentence (in natural language) such as

$$\text{Elephants are grey} \tag{1.2}$$

is a deterministic model which can be used to identify the race of an African mammal, on the basis of its colour. This model can be considered as a more elegant and intuitive expression of an equation such as

$$\text{colour}(\text{elephant}) = \text{grey},$$

or of the following formal rule:

$$\text{if animal}=\text{elephant then colour}=\text{grey}.$$

Also, if X_1 and X_2 are variables that correspond to the race and colour of a set of mammals, then the model can be converted in the formalism of this chapter:

$$\text{if } X_1 = \text{'elephant'}, \text{ then } X_2 = \text{'grey'}. \tag{1.3}$$

If the number of variables and the number of possible values of each of them are large, then the object can theoretically reside in a considerable number of states. Let us suppose however that all of these configurations can be enumerated and analyzed. Then the probabilistic modeling of the object consists in associating to any object state (or set of states), a *probability*, i.e., a number between 0 and 1, quantifying how plausible the object state (or set of states) is. We thus define the *joint probability distribution* of the set of variables X_1, X_2, \dots, X_n , denoted

$$\mathbb{P}(X_1, X_2, \dots, X_n).$$

The domain of this function is the Cartesian product $\mathcal{E}_1 \times \dots \times \mathcal{E}_n$ and its range is the interval $[0;1]$.

1.2.3 Benefits of probabilistic modeling

1.2.3.1 Modeling power

As far as modeling capability is concerned, probabilistic models are undeniably more powerful than deterministic ones. Indeed, a deterministic model may always be considered as a particular or simplified case of probabilistic model. For example, the model of sentence (1.2) above is a particular case of a slightly more complicated, probabilistic one:

$$x\% \text{ of elephants are grey} \quad (1.4)$$

where $x = 100$. This model can also be written using a conditional probability:

$$\mathbb{P}(X_2 = \text{'grey'} \mid X_1 = \text{'elephant'}) = x. \quad (1.5)$$

Incidentally, the probabilistic model is a more appropriate representation of reality in this example, since, for instance, a rare kind of elephant is white.

1.2.3.2 The power of doubt – exhaustiveness

Doubt is a typically human faculty which can be considered as the basis of any scientific process. This was also pointed out by Descartes, who recommended ‘never to accept anything for true which is not clearly known to be such; that is to say, carefully to avoid precipitancy and prejudice, and to comprise nothing more in one’s judgement than what was presented to one’s mind so clearly and distinctly as to exclude all ground of doubt.’ The construction of a probabilistic model requires the systematic examination of all possible values of each variable (each subset \mathcal{E}_j), and of each configuration of the object (i.e., each element of $\mathcal{E}_1 \times \cdots \times \mathcal{E}_n$). This reduces the impact of cultural and psychological biases and the risk to forget any important aspect of the object. Furthermore, it is hard to imagine a more precise representation of an object: each of the theoretically possible configurations of the object is considered, and to each of them is associated one element of the infinite set $[0;1]$.

1.2.3.3 Usability in a context of partial information

In many circumstances, probabilistic models are actually much easier to use than deterministic ones. Let us illustrate this with an example.

Example 3. A hiker has gone for a walk in a forest, and brings back home some flashy coloured mushrooms. He wants to decide whether he will have them for dinner, or not. Instead of consulting an encyclopedia of mushrooms, he phones a friend, with some knowledge of the domain. His friend tells him that:

75% of mushrooms with flashy colours are poisonous.

In this example, a deterministic model, such as an encyclopedia of mushrooms, would certainly help identify the exact nature of the mushrooms, but this requires an extensive examination, and takes some time. The probabilistic model provided by the hiker's friend is more suitable to satisfy his need, i.e., make a quick decision for his dinner, than the deterministic one. In fact, if the hiker wants to use the only available information 'the mushroom is flashy-coloured', then a form of probabilistic reasoning – possibly a very basic one – is absolutely necessary.

1.2.4 Drawbacks of probabilistic modeling

In spite of its benefits listed in the previous paragraph, the joint probability distribution $\mathbb{P}(X_1, X_2, \dots, X_n)$ is rarely employed *per se*. The reason is that this mathematical concept is rather unintuitive and difficult to handle.

Firstly, it can be graphically represented only if $n = 1$ or 2 . Even in the in the simplest nontrivial case $n = p = 2$ (illustrated in Figure 1.2), the graphical model is rather difficult to interpret. When $n \geq 3$, no graphical representation is possible, which, as mentioned above, restrains the model usability.

Secondly, the joint probability distribution gives rise to a phenomenon of combinatorial explosion. For instance, if each variable takes on p different values ($p \geq 1$), then the joint probability distribution has to be described by the probabilities of p^n potential configurations of the object, i.e., ten billion values if $n = p = 10$.

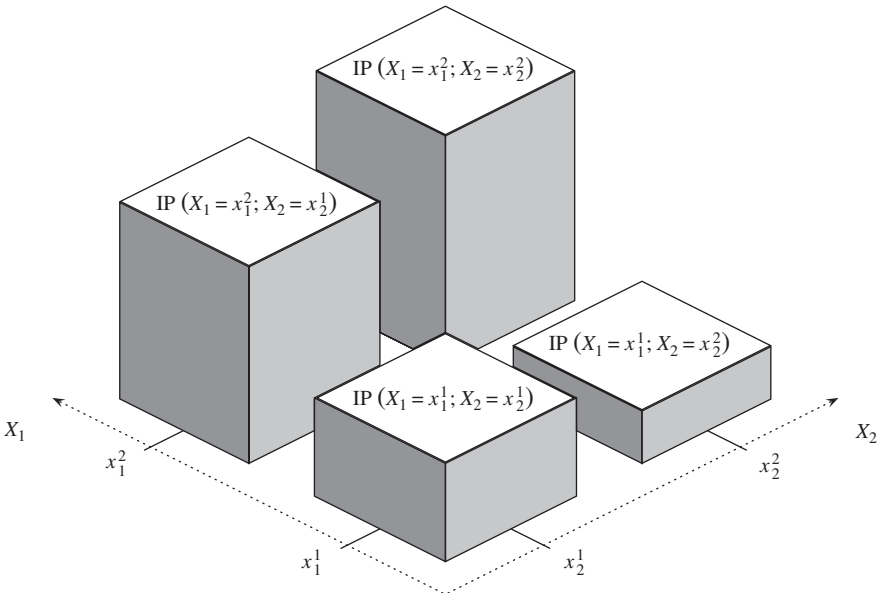


Figure 1.2 Representation of the joint probability distribution of a pair of random variables (X_1, X_2) .

1.3 Unconditional and conditional independence

1.3.1 Unconditional independence

Following Descartes's precept of dividing the difficulties, one may try to split the set of n variables into several subsets of smaller sizes which can relevantly be analyzed separately.

Suppose for example that the set of n variables may be divided into two subsets of sizes j and $n - j$ such as:

$$\mathbb{P}(X_1, X_2, \dots, X_n) = \mathbb{P}(X_1, \dots, X_j) \mathbb{P}(X_{j+1}, \dots, X_n). \quad (1.6)$$

Then the modeling problem can be transformed into two simpler ones. One can derive the joint probability of subset X_1, \dots, X_j , then that of subset X_{j+1}, \dots, X_n , and use Equation (1.6) to obtain the complete model.

The equality of two functions expressed by Equation (1.6) means that the subsets of variables (X_1, \dots, X_j) and (X_{j+1}, \dots, X_n) are independent, or – to avoid confusion with a concept which is defined below – *unconditionally* independent. This means that any information regarding the (X_1, \dots, X_j) subset (for instance, ' $X_1 = 7$ ' or ' $X_1 + X_2 > 3$ ') does not change the probability distribution of the second subset (X_{j+1}, \dots, X_n) .

However, unconditional independence between two subsets of variables is very unlikely to happen in real-world models. If it does happen, the initial definition of the object is not relevant: in such a case, it makes more sense to construct two separate models.

1.3.2 Conditional independence

A more common – or at least much more reasonably acceptable in real-world models – phenomenon is the so-called 'conditional independence'. Let us introduce this concept by two examples.

1.3.2.1 The lorry driver example

Example 4. A lorry driver is due to make a 600-mile trip. To analyze the risk of his falling asleep while driving, let us consider whether (1) he sleeps 'well' (more than seven hours) on the night before and (2) he feels tired at the beginning of the trip.

In this example, there are obvious causal relationships between the driver's sleep, his perceived fatigue, and the risk of falling asleep: the three variables are dependent. Let us suppose however that we know that the lorry driver feels tired at the beginning of the trip. Then knowing whether this is due to a bad sleep the previous night, or to any other reason is of no use to evaluate the risk. Similarly, if the lorry driver does *not* feel tired at the beginning of the trip, one may then consider that the quality of his sleep on the night before has no influence on the risk. Given these

considerations, the risk of falling asleep is said to be *conditionally independent* of the quality of sleep, *given* the lorry driver's fatigue.

To express it formally, let X_1 , X_2 and X_3 be binary variables telling whether the lorry driver sleeps well the night before, whether he feels tired at the beginning of the trip, and whether he will fall asleep while driving. Then X_3 is independent of X_1 , for any given value of X_2 . In terms of probabilities, we have:

$$\mathbb{P}(X_3 | X_1 \text{ and } X_2) = \mathbb{P}(X_3 | X_2). \quad (1.7)$$

In such a case, knowing the values of X_1 and X_2 is not better than knowing only the value of X_2 , and it is useless to describe the behavior of X_1, X_2, X_3 by a function of three variables; indeed, we may deduce from Equation (1.7):

$$\mathbb{P}(X_1, X_2, X_3) = \mathbb{P}(X_1) \mathbb{P}(X_2 | X_1) \mathbb{P}(X_3 | X_2), \quad (1.8)$$

which shows that the risk model can be constructed by successively studying the quality of sleep, then its influence on the state of fatigue, and then the influence of the state of fatigue on the risk of falling asleep.

1.3.2.2 The doped athlete example

Example 5. In a sports competition, each athlete undergoes two doping tests, aimed at detecting if he/she has taken a given prohibited substance: test A is a blood test and test B a urine test. The two tests are carried out in two different laboratories, without any form of consultation.

It is quite obvious in Example 5 that the results of the two tests are not independent variables. If test A is positive, then the participant is likely to have used the banned product; then test B will probably be also positive.

Now consider a participant who has taken the banned substance. Then tests A and B can be considered independent, since the two laboratories use different detection methods. Similarly, tests A and B can be considered independent when the participant has *not* taken the banned substance: the results of both tests are conditionally independent, given the status of the tested athlete. Formally, if X_1 is a binary variable telling whether the athlete is 'clean' or not, X_2 is the result of test A , and X_3 the result of test B , we can write:

$$\mathbb{P}(X_3 | X_1 \text{ and } X_2) = \mathbb{P}(X_3 | X_1). \quad (1.9)$$

Equation (1.9) can exactly be translated into 'knowing whether the athlete has taken the substance is enough information to estimate the chances of test B being positive'. A symmetrical equation holds regarding test A :

$$\mathbb{P}(X_2 | X_1 \text{ and } X_3) = \mathbb{P}(X_2 | X_1). \quad (1.10)$$

Here again, it is useless to describe the behavior of X_1, X_2, X_3 by a function of three variables. Equations (1.9) and (1.10) yield:

$$\mathbb{P}(X_1, X_2, X_3) = \mathbb{P}(X_1) \mathbb{P}(X_2 | X_1) \mathbb{P}(X_3 | X_1), \quad (1.11)$$

which means that considerations on the proportion of doped athletes $\mathbb{P}(X_1)$, and on the reliabilities of each tests are sufficient to construct the model.

1.4 Bayesian networks

1.4.1 Examples

In the lorry driver and doped athlete examples, we have identified the most direct and significant influences between the variables, and simplified the derivation of the joint probability distribution. By representing these influences in a graphical form, we now introduce the notion of Bayesian network.

In Example 4, our analysis has shown that there is an influence of variable X_1 on variable X_2 , and another influence of variable X_2 on variable X_3 ; we have assumed that there is no direct relation between X_1 and X_3 . The usual way of representing such influences is a diagram of nodes and arrows, connecting influencing variables (parent variables) to influenced variables (child variables). The structure corresponding to Example 4 is shown in Figure 1.3.

Similarly, the influences analyzed in Example 5 may be represented as shown in Figure 1.4.

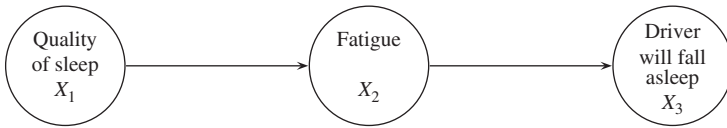


Figure 1.3 A representation of the influences between variables in Example 4. Variable X_3 is *conditionally independent* of X_1 given X_2 .

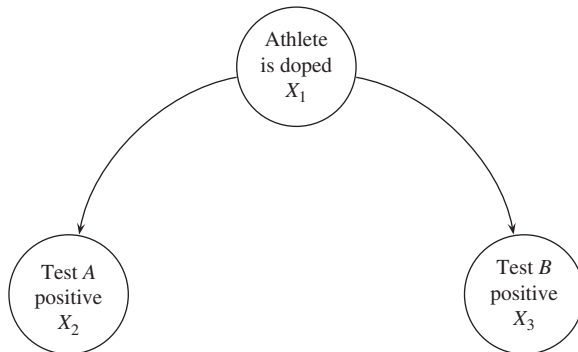


Figure 1.4 A representation of the influences between variables in Example 5. Variables X_2 and X_3 are *conditionally independent* given X_1 .

Considering the graphical structures of Figures 1.3 and 1.4, and more precisely the parents of each variable, we observe that both Equations (1.8) and (1.11) can be written in the following form:

$$\mathbb{P}(X_1, X_2, X_3) = \mathbb{P}(X_1 | \text{parents}(X_1)) \mathbb{P}(X_2 | \text{parents}(X_2)) \mathbb{P}(X_3 | \text{parents}(X_3)). \quad (1.12)$$

Equation (1.12) is the formal definition of a Bayesian network, in the three-variable case: through a process of analyzing and sorting out the unconditional independences between the three variables, we have been able to convert $\mathbb{P}(X_1, X_2, X_3)$ into a product of three conditional probabilities. This definition is generalized in the next paragraph.

1.4.2 Definition

Definition 6 (Bayesian network) *Let us consider n random variables X_1, X_2, \dots, X_n , a directed acyclic graph with n numbered nodes, and suppose node j ($1 \leq j \leq n$) of the graph is associated to the X_j variable. Then the graph is a Bayesian network, representing the variables X_1, X_2, \dots, X_n , if:*

$$\mathbb{P}(X_1, X_2, \dots, X_n) = \prod_{j=1}^n \mathbb{P}(X_j | \text{parents}(X_j)), \quad (1.13)$$

where: $\text{parents}(X_j)$ denotes the set of all variables X_i , such that there is an arc from node i to node j in the graph.

As shown in the examples, Equation (1.13) simplifies the calculation of the joint probability distribution. Let us suppose for instance that each variable has p possible values, and less than three parent variables. Then the number of probabilities in the model is lower than $n \cdot p^4$, although the object can reside in p^n configurations. If $n = p = 10$, the reduction factor is greater than one hundred thousands.

A crucial point is that this simplification is based on an graphical, intuitive representation, and not on some highly technical considerations. A diagram of boxes and arrows can be easily interpreted, discussed and validated on a step-by-step basis by the stakeholders: there is no ‘black box’ effect in the modeling process.

Another important remark can be deduced from Definition 6.

Proposition 7. *Any joint probability distribution may be represented by a Bayesian network.*

Indeed, we may formally express $\mathbb{P}(X_1, X_2, \dots, X_n)$ as follows:

$$\begin{aligned} \mathbb{P}(X_1, X_2, \dots, X_n) &= \mathbb{P}(X_1) \mathbb{P}(X_2, \dots, X_n | X_1) \\ &= \mathbb{P}(X_1) \mathbb{P}(X_2 | X_1) \cdots \mathbb{P}(X_3, \dots, X_n | X_1, X_2) \\ &= \dots \\ &= \mathbb{P}(X_1) \mathbb{P}(X_2 | X_1) \cdots \mathbb{P}(X_n | X_1, \dots, X_{n-1}). \end{aligned} \quad (1.14)$$

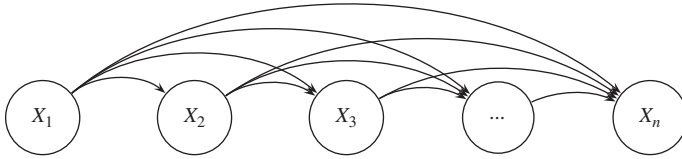


Figure 1.5 ‘Generic’ structure of a Bayesian Network, suitable for any joint probability distribution of n random variables X_1, \dots, X_n .

Equation (1.14) proves that the structure of Figure 1.5, with one arc from variable X_i to X_j , whenever $i < j$, is suitable to represent any joint probability distribution between n variables. In other words, there is no loss of generality in modeling a set of random variables with a Bayesian network.

Proposition 7 helps to answer a frequently asked question: *in Definition 6, why has the graph to be acyclic?* Besides the fact that Equation (1.13) would not make sense in the presence of loops, the hypothesis of graphical acyclicity is not at all restrictive: whatever the number and nature of the dependencies between the X_i variables, there is always at least one acyclic structure (i.e., that of Figure 1.5) that is suitable to represent the object.

Proposition 7 clearly shows the modeling power of Bayesian Networks. As mentioned above, any deterministic model is a particular case of probabilistic model; any probabilistic model may be represented as a Bayesian network.

