Chapter 10

Graphical models

The previous chapters discussed models that relate the observed measurements to some aspect of the world that we wish to estimate. In each case, this relationship depended on a set of parameters and for each model we presented a learning algorithm that estimated these parameters.

Unfortunately, the utility of these models is limited because every element of the model depends on every other. For example, in generative models we model the joint probability of the observations and the world state. In many problems both of these quantities may be high dimensional. Consequently, the number of parameters required to characterize their joint density accurately is very large. Discriminative models suffer from the same pathology: if every element of the world state depends on every element of the data, a large number of parameters will be required to characterize this relationship. In practice, the required amount of training data and the computational burden of learning and inference reach impractical levels.

The solution to this problem is to reduce the dependencies between variables in the model by identifying (or asserting) some degree of redundancy. To this end, we introduce the idea of *conditional independence*, which is a way of characterizing these redundancies. We then introduce *graphical models* which are graph-based representations of the conditional independence relations. We discuss two different types of graphical models — directed and undirected — and we consider the implications for learning, inference, and drawing samples.

This chapter does not develop specific models or discuss vision applications. The goal is to provide the theoretical background for the models in subsequent chapters. We will illustrate the ideas with probability distributions where the constituent variables are discrete; however, almost all of the ideas transfer directly to the continuous case.

10.1 Conditional independence

When we first discussed probability distributions, we introduced the notion of independence (section 2.6). Two variables x_1 and x_2 are independent if their joint

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

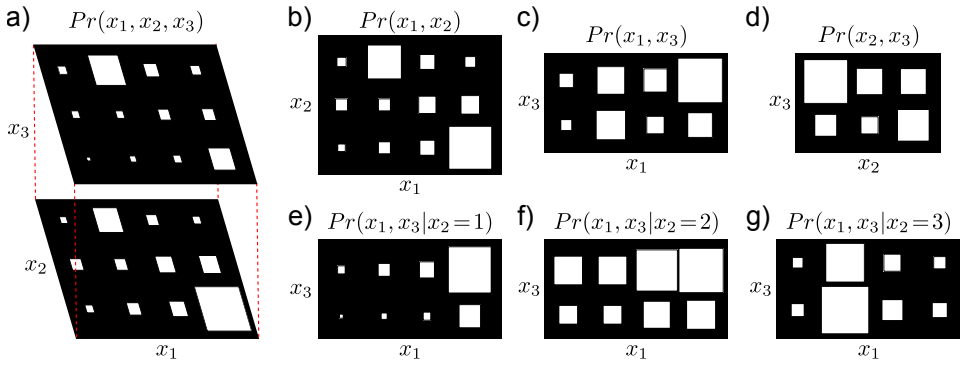

Figure 10.1 Conditional independence. a) Joint pdf of three discrete variables x_1, x_2, x_3, which take four, three, and two possible values, respectively. All 24 probability values sum to one. b) Marginalizing, we see that variables x_1 and x_2 are dependent; the conditional distribution of x_1 is different for different values of x_2 (the elements in each row are not in the same proportions)and vice-versa. c) Variables x_1 and x_3 are also dependent. d) Variables x_2 and x_3 are also dependent. e-g) However, x_1 and x_3 are conditionally independent *given* x_2. For fixed x_2, x_1 tells us nothing more about x_3 and vice-versa.

probability distribution factorizes as $Pr(x_1, x_2) = Pr(x_1)Pr(x_2)$. In layman's terms, one variable provides no information about the other if they are independent.

With more than two random variables, independence relations become more complex. The variable x_1 is said to be *conditionally independent of variable* x_3 *given variable* x_2 when x_1 and x_3 are independent for fixed x_2 (figure 10.1). In mathematical terms, we have

$$\begin{aligned} Pr(x_1|x_2, x_3) &= Pr(x_1|x_2) \\ Pr(x_3|x_1, x_2) &= Pr(x_3|x_2). \end{aligned} \tag{10.1}$$

Note that conditional independence relations are always symmetric; if x_1 is conditionally independent of x_3 given x_2, then it is also true that x_3 is independent of x_1 given x_2.

Confusingly, the conditional independence of x_1 and x_3 given x_2 does not mean that x_1 and x_3 are themselves independent. It merely implies that if we know variable x_2 then x_1 provides no further information about x_3 and vice-versa. One way that this can occur is in a chain of events: if event x_1 causes event x_2 and x_2 causes x_3 then the dependence of x_3 on x_1 might be entirely mediated by x_2.

Now consider decomposing the joint probability distribution $Pr(x_1, x_2, x_3)$ into the product of conditional probabilities. When x_1 is independent of x_3 given x_2, we find that

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

$$\begin{aligned} Pr(x_1, x_2, x_3) &= Pr(x_3|x_2, x_1)Pr(x_2|x_1)Pr(x_1) \\ &= Pr(x_3|x_2)Pr(x_2|x_1)Pr(x_1). \end{aligned} \tag{10.2}$$

The conditional independence relation means that the probability distribution factorizes in a certain way (and is hence redundant). This redundancy implies that we can describe the distribution with fewer parameters and so working with models with large numbers of variables becomes more tractable.

Throughout this chapter, we will explore the relationship between factorization of the distribution and conditional independence relations. To this end, we will introduce graphical models. These are graph-based representations that make both the factorization and the conditional independence relations easy to establish. In this book we will consider two different types of graphical model – directed and undirected graphical models – each of which corresponds to a different type of factorization.

10.2 Directed graphical models

A *directed graphical model* or *Bayesian network* represents the factorization of the joint probability distribution into a product of conditional distributions that take the form of a directed acyclic graph (DAG) so that

$$Pr(x_{1\ldots N}) = \prod_{n=1}^{N} Pr(x_n|x_{\mathrm{pa}[n]}), \tag{10.3}$$

where $\{x_n\}_{n=1}^N$ represent the constituent variables of the joint distribution and the function $\mathrm{pa}[n]$ returns the indices of variables that are parents of variable x_n.

We can visualize the factorization as a directed graphical model (figure 10.2) by adding one node per random variable and drawing an arrow to each variable x_n from each of its parents $x_{\mathrm{pa}[n]}$. This directed graphical model should never contain cycles. If it does, then the original factorization was not a valid probability distribution.

Problem 10.1
Problem 10.2

To retrieve the factorization from the graphical model, we introduce one factorization term per variable in the graph. If variable x_n is independent of all others (has no parents), then we write $Pr(x_n)$. Otherwise, we write $Pr(x_n|x_{\mathrm{pa}[n]})$ where the parents $x_{\mathrm{pa}[n]}$ consist of the set of variables with arrows that point to x_n.

10.2.1 Example 1

The graphical model in figure 10.2 represents the factorization

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

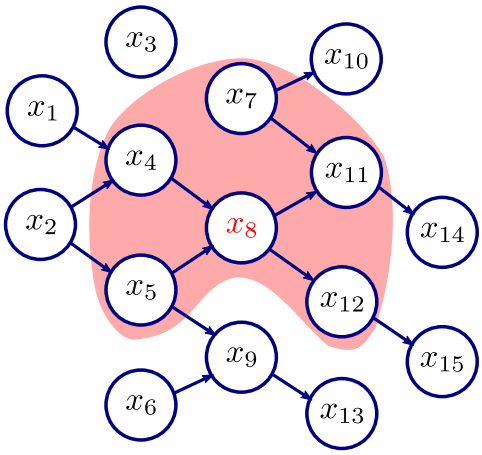

Figure 10.2 Example 1. A directed graphical model has one node per term in the factorization of the joint probability distribution. A node x_n with no incoming connections represents the term $Pr(x_n)$. A node x_n with incoming connections $\mathbf{x}_{\mathrm{pa}[n]}$ represents the term $Pr(x_n|x_{\mathrm{pa}[n]})$. Variable x_n is conditionally independent of all of the others given its *Markov blanket*. This comprises its parents, its children and other parents of its children. For example, the Markov blanket for variable x_8 is indicated by the shaded region.

$$\begin{aligned} Pr(x_1 \ldots x_{15}) \;=\; & Pr(x_1)Pr(x_2)Pr(x_3)Pr(x_4|x_1,x_2)Pr(x_5|x_2)Pr(x_6) \\ & Pr(x_7)Pr(x_8|x_4,x_5)Pr(x_9|x_5,x_6)Pr(x_{10}|x_7)Pr(x_{11}|x_7,x_8) \\ & Pr(x_{12}|x_8)Pr(x_{13}|x_9)Pr(x_{14}|x_{11})Pr(x_{15}|x_{12}). \end{aligned} \tag{10.4}$$

The graphical model (or factorization) implies a set of independence and conditional independence relations between the variables. Some statements about these relations can be made based on a superficial look at the graph. First, if there is no directed path between two variables following the arrow directions and they have no common ancestors, then they are independent. So, variable x_3 in figure 10.2 is independent of all of the other variables, and variables x_1 and x_2 are independent of each other. Variables x_4 and x_5 are not independent as they share an ancestor. Second, any variable is conditionally independent of all the other variables given its parents, children, and the other parents of its children (its *Markov blanket*). So, for example, variable x_8 in figure 10.2 is conditionally independent of the remaining variables given those in the shaded area.

For vision applications, these rules are usually sufficient to gain an understanding of the main properties of a graphical model. However, occasionally we may wish to test whether one arbitrary set of nodes is independent of another given a third. This is not easily established by looking at the graph, but can be tested using the following criterion:

The variables in set $\mathcal{A}$ are conditionally independent of those in set $\mathcal{B}$ given set $\mathcal{C}$ if all routes from $\mathcal{A}$ to $\mathcal{B}$ are blocked. A route is blocked at a node if (i) this node is in $\mathcal{C}$ and the arrows meet head to tail or tail to tail or (ii) neither this node nor any of its descendants are in $\mathcal{C}$ and the arrows meet head to head.

See Koller & Friedman 2009 for more details of why this is the case.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

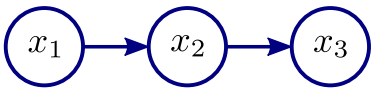

Figure 10.3 Example 2. Directed graphical model relating variables x_1, x_2, x_3 from figure 10.1. This model implies that the joint probability can be broken down as $Pr(x_1, x_2, x_3) = Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_2)$.

10.2.2 Example 2

Figure 10.3 tells us that

$$Pr(x_1, x_2, x_3) = Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_2). \tag{10.5}$$

In other words, this is the graphical model corresponding to the distribution in figure 10.1.

If we condition on x_2, then the only route from x_1 to x_3 is blocked at x_2 (the arrows meet head to tail here) and so x_1 must be conditionally independent of x_3 given x_2. We could have reached the same conclusion by noticing that the Markov blanket for variable x_1 is just variable x_2.

In this case, it is easy to prove this conditional independence relation algebraically. Writing out the conditional probability of x_1 given x_2 and x_3

$$\begin{aligned} Pr(x_1|x_2, x_3) &= \frac{Pr(x_1, x_2, x_3)}{Pr(x_2, x_3)} \\ &= \frac{Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_2)}{\int Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_2)dx_1} \\ &= \frac{Pr(x_1)Pr(x_2|x_1)}{\int Pr(x_1)Pr(x_2|x_1)dx_1}, \end{aligned} \tag{10.6}$$

we see that the final expression does not depend on x_3 and so we deduce that x_1 is conditionally independent of x_3 given x_2 as required.

Notice that the factorized distribution is more efficient to represent than the full version. The original distribution $Pr(x_1, x_2, x_3)$ (figure 10.1a) contains $4{\times}3{\times}2{=}24$ entries. However, the terms $Pr(x_1)$, $Pr(x_2|x_1)$, and $Pr(x_3|x_2)$ contain 4, $4{\times}3{=}12$, and $3 \times 2 = 6$ entries, respectively, giving a total of 22 entries. In this case, this is not a dramatic reduction, but in more practical situations it would be. For example, if each variable took 10 possible values, the full joint distribution would have $10 \times 10 \times 10 = 1000$ values, but the factorized distribution would have only $10{+}100{+}100{=}210$ values. For even larger systems, this can make a huge saving. One way to think about conditional independence relations is to consider them as redundancies in the full joint probability distribution.

10.2.3 Example 3

Finally, in figure 10.4 we present graphical models for the mixture of Gaussians, t-distribution and factor analysis models from chapter 7. These depictions imme-

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

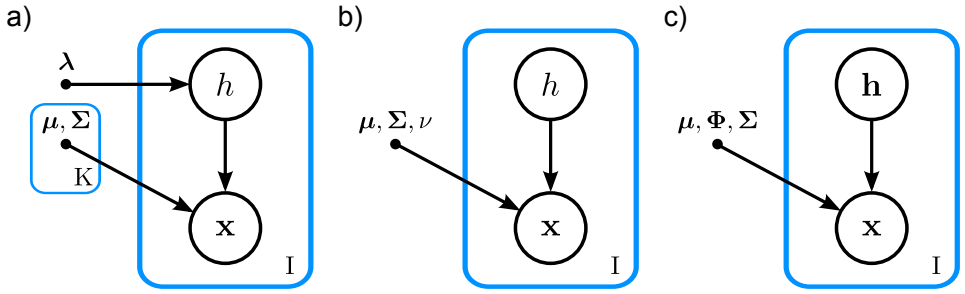

Figure 10.4 Example 3. Graphical models for a) mixture of Gaussians b) t-distribution and c) factor analysis. A node (black circle) represents a random variable. In a graphical model a bullet • represents a variable whose value is considered to be fixed. Each variable may be repeated many times, and this is indicated by a plate (blue rectangle) where the number of copies is indicated in the lower right corner. For example, in a) there are I copies of the training data $\{\mathbf{x}_i\}_{i=1}^{I}$ and I copies of the variable $\{h_i\}_{i=1}^{I}$. Similarly, there are K sets of parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$, but just one weight vector $\boldsymbol{\lambda}$.

diately demonstrate that these models have very similar structures.

They also add several new features to the graphical representation. First, they include multidimensional variables. Second, they include variables that are considered as fixed and these are marked by a bullet •. We condition on the fixed variables, but do not define a probability distribution over them. Figure 10.4c depicts the factorization $Pr(\mathbf{h}_i, \mathbf{x}_i) = Pr(\mathbf{h}_i)Pr(\mathbf{x}_i|\mathbf{h}_i, \boldsymbol{\mu}, \boldsymbol{\Phi}, \boldsymbol{\Sigma})$.

Finally, we have also used *plate* notation. A plate is depicted as a rectangle with a number in the corner. It indicates that the quantities inside the rectangle should be repeated the given number of times. For example, in figure 10.4c there are I copies $\{\mathbf{x}_i, \mathbf{h}_i\}_{i=1}^{I}$ of the variables $\mathbf{x}$ and $\mathbf{h}$ but only one set of parameters $\boldsymbol{\mu}, \boldsymbol{\Phi}$, and $\boldsymbol{\Sigma}$.

10.2.4 Summary

To summarize, we can think about the structure of the joint probability distribution in three ways. First, we can consider the way that the probability distribution factorizes. Second, we can examine the directed graphical model. Third, we can think about the conditional independence relations.

There is a one-to-one mapping between directed graphical models (acyclic directed graphs of conditional probability relations) and factorizations. However, the relationship between the graphical model (or factorization) and the conditional independence relations is more complicated. A directed graphical model (or its equivalent factorization) determines a set of conditional independence relations. However, as we shall see later in this chapter, there are some sets of conditional independence relations that cannot be represented by directed graphical models.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

10.3 Undirected graphical models

In this section we introduce a second family of graphical models. Undirected graphical models represent probability distributions over variables $\{x_n\}_{n=1}^N$ that take the form of a product of *potential functions* $\phi[x_{1\ldots N}]$ so that

Problem 10.3
Problem 10.4
Problem 10.5

$$Pr(x_{1\ldots N}) = \frac{1}{Z}\prod_{c=1}^{C}\phi_c[x_{1\ldots N}], \tag{10.7}$$

where the potential function $\phi_c[x_{1\ldots N}]$ always returns a positive number. Since the probability increases when $\phi_c[x_{1\ldots N}]$ increases, each of these functions modulates the tendency for the variables $x_{1\ldots N}$ to take certain values. The probability is greatest where all of the functions $\phi_{1\ldots C}$ return high values. However, it should be emphasized that potential functions are *not* the same as conditional probabilities, and there is not usually a clear way to map from one to the other.

The term Z is known as the *partition function* and normalizes the product of these positive functions so that the total probability is one. In the discrete case, it would be computed as

$$Z = \sum_{x_1}\sum_{x_2}\cdots\sum_{x_N}\prod_{c=1}^{C}\phi_c[x_{1\ldots N}]. \tag{10.8}$$

For realistically sized systems, this sum will be intractable; we will not be able to compute Z and hence will only be able to compute the overall probability up to an unknown scale factor.

We can equivalently write equation 10.7 as

$$Pr(x_{1\ldots N}) = \frac{1}{Z}\exp\left[-\sum_{c=1}^{C}\psi_c[x_{1\ldots N}]\right], \tag{10.9}$$

where $\psi_c[x_{1\ldots N}] = -\log[\phi_c[x_{1\ldots N}]]$. When written in this form, the probability is referred to as a *Gibbs distribution*. The terms $\psi_c[x_{1\ldots N}]$ are functions that may return any real number and can be thought of as representing a cost for every combination of labels $x_{1\ldots N}$. As the cost increases, the probability decreases. The total cost $\sum_{c=1}^{C}\psi_c[x_{1\ldots N}]$ is sometimes known as the *energy*, and the process of fitting the model (increasing the probability) is hence sometimes termed *energy minimization*.

When each potential function $\phi[\bullet]$ (or alternatively each cost function $\psi[\bullet]$) addresses all of the variables $x_{1\ldots N}$ the undirected graphical model is known as a *product of experts*. However, in computer vision it is more common for each potential function to operate on a subset of the variables $\mathcal{S} \subset \{x_n\}_{n=1}^N$. These subsets are called *cliques* and it is the choice of these cliques that determines the conditional independence relations. Denoting the c^{th} clique by $\mathcal{S}_c$ we can rewrite equation 10.7 as

$$Pr(x_{1\ldots N}) = \frac{1}{Z}\prod_{c=1}^{C}\phi_c[\mathcal{S}_c]. \tag{10.10}$$

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

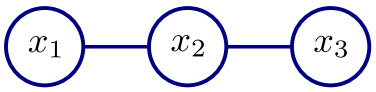

Figure 10.5 Example 1. Undirected graphical model relating variables x_1, x_2, and x_3. This model implies that the joint probability can be factorized as $Pr(x_1, x_2, x_3) = \frac{1}{Z}\phi_1[x_1, x_2]\phi_2[x_2, x_3]$.

In other words, the probability distribution is factorized into a product of terms, each of which only depends on a subset of variables. In this situation, the model is sometimes referred to as a *Markov random field*.

To visualize the undirected graphical model, we draw one node per random variable. Then, for every clique $\mathcal{S}_c$ we draw a connection from every member variable $x_i \in \mathcal{S}_c$ to every other member variable.

Moving in the opposite direction, we can take a graphical model and establish the underlying factorization using the following method. We add one term to the factorization per *maximal clique* (see figure 10.6). A maximal clique is a fully connected subset of nodes (i.e., a subset where every node is connected to every other) where it is not possible to add another node and remain fully connected.

It is much easier to establish the conditional independence relations from an undirected graphical model than for directed graphical models. They can be found using the following property:

One set of nodes is conditionally independent of another given a third if the third set separates them (prevents a path from the first node to the second).

It follows that a node is conditionally independent of all other nodes given its set of immediate neighbors, and so these neighbors form the Markov blanket.

10.3.1 Example 1

Consider the graphical model in figure 10.5. This represents the factorization

$$Pr(x_1, x_2, x_3) = \frac{1}{Z}\phi_1[x_1, x_2]\phi_2[x_2, x_3]. \tag{10.11}$$

We can immediately see that variable x_1 is conditionally independent of variable x_3 given x_2 because x_2 separates the other two variables: it blocks the path from x_1 to x_3. In this case, the conditional independence relation is easy to prove:

$$\begin{aligned} Pr(x_1|x_2, x_3) &= \frac{Pr(x_1, x_2, x_3)}{Pr(x_2, x_3)} \\ &= \frac{\frac{1}{Z}\phi_1[x_1, x_2]\phi_2[x_2, x_3]}{\int \frac{1}{Z}\phi_1[x_1, x_2]\phi_2[x_2, x_3]dx_1} \\ &= \frac{\phi_1[x_1, x_2]}{\int \phi_1[x_1, x_2]dx_1}. \end{aligned} \tag{10.12}$$

The final expression does not depend on x_3 and so we conclude that x_1 is conditionally independent of x_3 given x_2.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

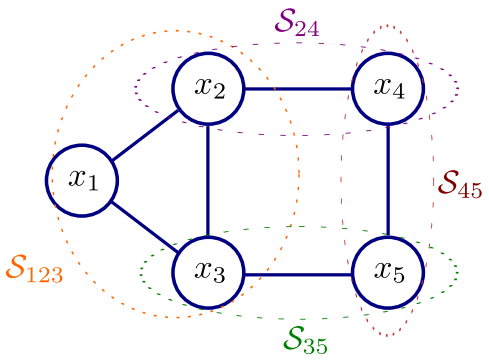

Figure 10.6 Example 2. Undirected graphical model representing variables $\{x_i\}_{i=1}^{5}$. The associated probability distribution factorizes into a product of one potential function per maximal clique. The clique $\mathcal{S}_{45} = \{x_4, x_5\}$ is a maximal clique as there is no other node that we can add that connects to every node in the clique. The clique $\mathcal{S}_{23} = \{x_2, x_3\}$ is not a maximal clique as it is possible to add node x_1 and all three nodes in the new clique are connected to each other.

10.3.2 Example 2

Consider the graphical model in figure 10.6. There are four maximal cliques in this graph, and so it represents the factorization

$$Pr(x_{1\ldots5}) = \frac{1}{Z}\phi_1[x_1, x_2, x_3]\phi_2[x_2, x_4]\phi_3[x_3, x_5]\phi_4[x_4, x_5]. \tag{10.13}$$

We can deduce various conditional independence relations from the graphical representation. For example, variable x_1 is conditionally independent of variables x_4 and x_5 given x_2 and x_3, and variable x_5 is independent of variables x_1 and x_2 given x_3 and x_4, and so on.

Note also that the factorization

$$Pr(x_{1\ldots5}) = \frac{1}{Z}\left(\phi_1[x_1, x_2]\phi_2[x_2, x_3]\phi_3[x_1, x_3]\right)\phi_4[x_2, x_4]\phi_5[x_3, x_5]\phi_6[x_4, x_5] \tag{10.14}$$

creates the same graphical model; there is a many-to-one mapping from factorizations to undirected graphical models (as opposed to the one-to-one mapping for directed graphical models). When we compute a factorization from the graphical model based on the maximal cliques we do so in a conservative way. It is possible that there are further redundancies which were not made explicit by the undirected graphical model.

10.4 Comparing directed and undirected graphical models

In sections 10.2 and 10.3 we have discussed directed and undirected graphical models, respectively. Each graphical model represents a factorization of the probability distribution. We have presented methods to extract the conditional independence relations from each type of graphical model. The purpose of this section is to argue that these representations are not equivalent. There are patterns of conditional independence that can be represented by directed graphical models but not undirected graphical models and vice versa.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

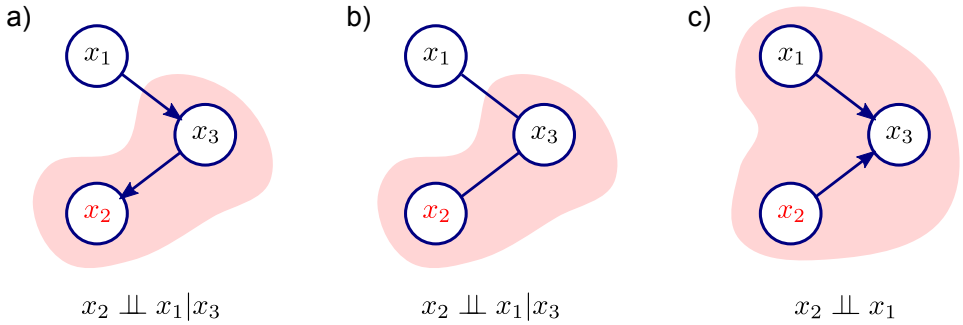

Figure 10.7 Directed vs. undirected graphical models. a) Directed graphical model with three nodes. There is only one conditional independence relation implied by this model: the node x_3 is the Markov blanket of node x_2 (shaded area) and so $x_2 \perp\!\!\!\perp x_1 | x_3$, where the notation $\perp\!\!\!\perp$ can be read as 'is independent of.' b) This undirected graphical model implies the same conditional independence relation. c) Second directed graphical model. The relation $x_2 \perp\!\!\!\perp x_1 | x_3$ is no longer true, but x_1 and x_2 are independent if we don't condition on x_3 so we can write $x_2 \perp\!\!\!\perp x_1$. There is no undirected graphical model with three variables that has this pattern of independence and conditional independence.

Problem 10.6
Problem 10.7
Problem 10.8

Figures 10.7a-b show an undirected and directed graphical model that do represent the same conditional independence relations. However, figure 10.7c shows a directed graphical model for which there is no equivalent undirected graphical model. There is simply no way to induce the same pattern of independence and conditional independence with an undirected graphical model.

Conversely, figure 10.8a shows an undirected graphical model that induces a pattern of conditional independence relations that cannot be replicated by any directed graphical model. Figure 10.8b shows a directed graphical model that is close, but still not equivalent; the Markov blanket of x_2 is different in each model and so are its conditional independence relations.

We conclude from this brief argument that directed and undirected graphical models do not represent the same subset of independence and conditional independence relations, and so we cannot eliminate one or the other from our consideration. In fact, there are other patterns of conditional independence that cannot be represented by either type of model. However, these will not be considered in this book. For further information concerning the families of distributions that can be represented by different types of graphical model, consult Barber (2012) or Koller & Friedman (2009).

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

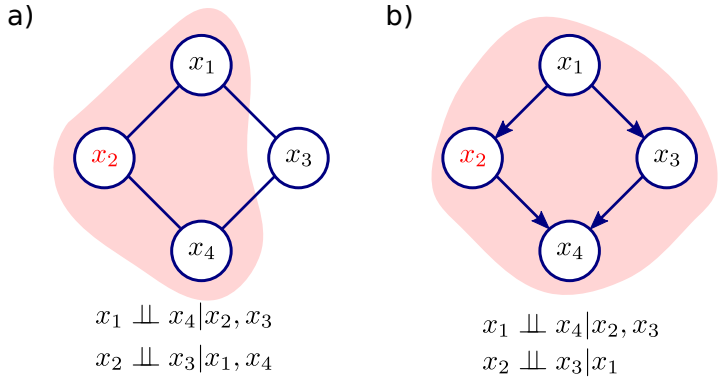

Figure 10.8 Directed vs. undirected models. a) This undirected graphical model induces two conditional independence relations. However, there is no equivalent directed graphical model that produces the same pattern. b) This directed graphical model also induces two conditional independence relations but they are not the same. In both cases, the shaded region represents the Markov blanket of variable x_2.

10.5 Graphical models in computer vision

We will now introduce a number of common vision models and look at their associated graphical models. We will discuss each of these in detail in subsequent chapters. However, it is instructive to see them together.

Problem 10.9
Problem 10.10

Figure 10.9a shows the graphical model for a *hidden Markov model* or *HMM*. We observe a sequence of measurements $\{\mathbf{x}_n\}_{n=1}^N$ each of which tells us something about the corresponding discrete world state $\{\mathbf{w}_n\}_{n=1}^N$. Adjacent world states are connected together so that the previous world state influences the current one and potentially resolves situations where the measurements are ambiguous. A prototypical application would be tracking sequences of sign language gestures (figure 10.9b). There is information at each frame about which gesture is present, but it may be ambiguous. However, we can impose prior knowledge that certain signs are more likely to follow others using the HMM and get an improved result.

Figure 10.9c represents a *Markov tree*. Again we observe a number of measurements, each of which provides information about the associated discrete world state. However, the world states are now connected in a tree structure. A prototypical application would be human body fitting (figure 10.9d) where each unknown world state represents a body part. The parts of the body naturally have a tree structure and so it makes sense to build a model that exploits this.

Figure 10.9e illustrates the use of a *Markov random field* or *MRF* as a prior. The MRF here describes the world state as a grid of undirected connections. Each node might correspond to a pixel. There is also a measurement variable associated with each world state variable. These pairs are connected with directed links, so overall this is a mixed model (partly directed and partly undirected). A prototypical application of an MRF in vision would be for semantic labeling (figure 10.9f). The

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

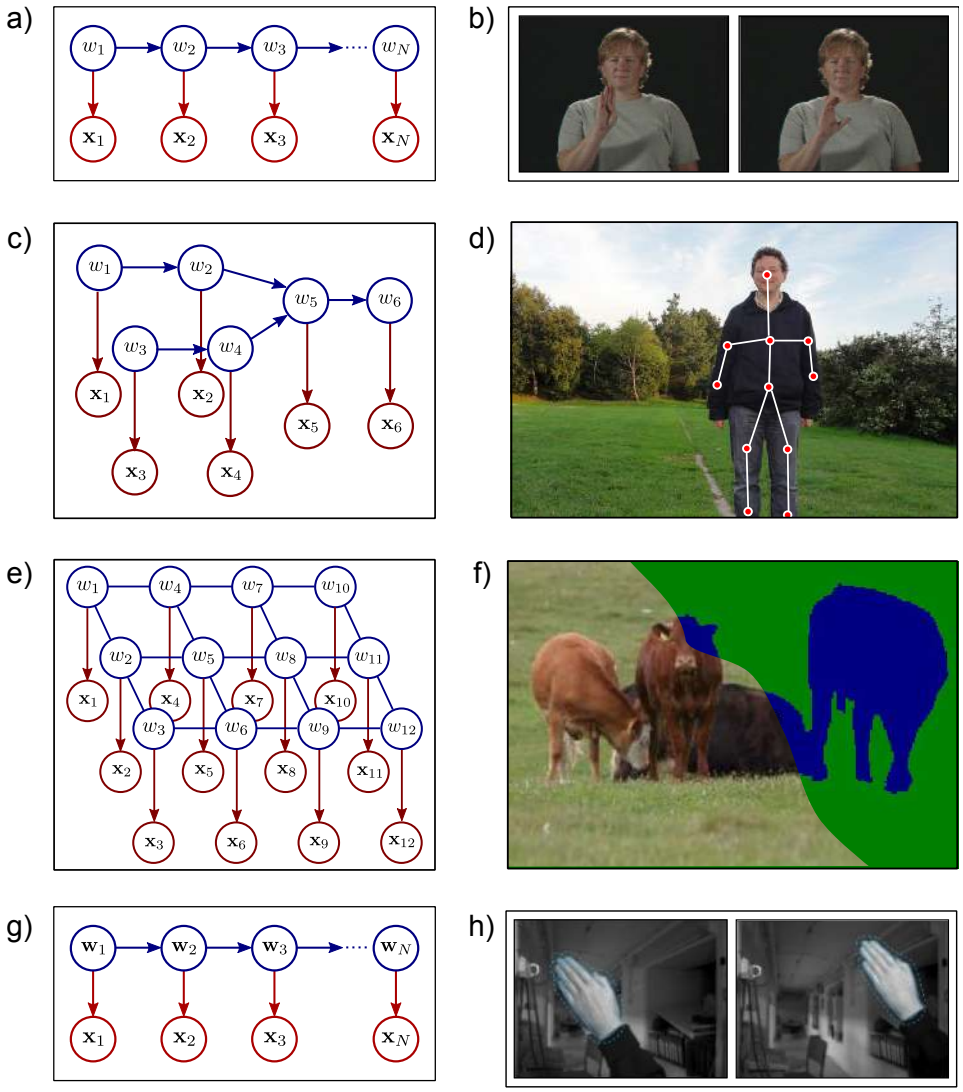

Figure 10.9 Commonly used graphical models in computer vision. a) Hidden Markov model (HMM). b) One possible application of the HMM is interpreting sign language sequences. The choice of sign at time n depends on the sign at time $n-1$. c) Markov tree. d) An example application is fitting a tree-structured body model e) Markov random field (MRF) prior with independent observations. f) The MRF is often used as a prior distribution in semantic labeling tasks. Here the goal is to infer a binary label at each pixel determining whether it belongs to the cow or the grass. g) Kalman filter. An example application is tracking an object through a sequence. It has the same graphical model as the HMM, but the unknown quantities are continuous as opposed to discrete.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

measurements constitute the RGB values at each position. The world state at each pixel is a discrete variable that determines the class of object present (i.e., cow vs. grass). The Markov random field prior ties together all of the individual classifiers to help yield a solution that makes global sense.

Finally, figure 10.9g shows the *Kalman filter*. This has the same graphical model as the hidden Markov model but in this case the world state is continuous rather than discrete. A prototypical application of the Kalman filter is for tracking objects through a time sequence (figure 10.9h). At each time, we might want to know the 2D position, size, and orientation of the hand. However, in a given frame the measurements might be poor: the frame may be blurred or the object may be temporarily occluded. By building a model that connects the estimates from adjacent frames, we can increase the robustness to these factors; earlier frames can resolve the uncertainty in the current ambiguous frame.

Notice that all of these graphical models have directed links from the world $\mathbf{w}$ to the data $\mathbf{x}$ that indicate a relationship of the form $Pr(\mathbf{x}|\mathbf{w})$. Hence, they all construct a probability distribution over the data and are generative models. We will also consider discriminative models, but, historically speaking, generative models of this kind have been more important. Each model is quite sparsely connected: each data variable $\mathbf{x}$ connects only to one world state variable $\mathbf{w}$, and each world state variable connects to only a few others. The result of this is that there are many conditional independence relations in the model. We will exploit these redundancies to develop efficient algorithms for learning and inference.

We will return to all of these models later in the book. We investigate the hidden Markov model and the Markov tree in chapter 11. We discuss the Markov random field in chapter 12, and we will present the Kalman filter in chapter 19. The remaining part of this chapter answers two questions: (i) How can we perform inference when there are a large number of unknown world variables? (ii) What are the implications of using a directed graphical model vs. an undirected one?

10.6 Inference in models with many unknowns

We will now consider inference in these models. Ideally, we would compute the full posterior distribution $Pr(w_{1\ldots N}|\mathbf{x}_{1\ldots N})$ using Bayes' rule. However, the unknown world states in the preceding models are generally much larger than previously considered in this book and this makes inference challenging.

For example, consider the space of world states in the HMM example. If we are given 1000 frames of video and there are 500 common signs in the sign language, then there are 500^{1000} possible states. It is clearly not practical to compute and store the posterior probability associated with each. Even when the world states are continuous, computing and storing the parameters of a high-dimensional probability model is still problematic. Fortunately, there are alternative approaches to inference, which we now consider in turn.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

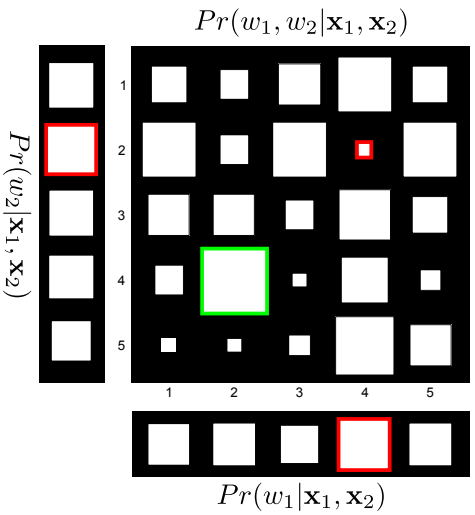

Figure 10.10 MAP solution vs. max marginals solution. The main figure shows the joint posterior distribution $Pr(w_1, w_2|\mathbf{x}_1, \mathbf{x}_2)$. The MAP solution is at the peak of this distribution at $w_1 = 2, w_2 = 4$ (highlighted in green). The figure also shows the two marginal distributions $Pr(w_1|\mathbf{x}_1, \mathbf{x}_2)$ and $Pr(w_2|\mathbf{x}_1, \mathbf{x}_2)$. The maximum marginals solution is computed by individually finding the maximum of each marginal distributions, which gives the solution $w_1 = 4, w_2 = 2$ (highlighted in red). For this distribution, this is very unrepresentative; although these labels are individually likely, they rarely co-occur and the joint posterior for this combination has low probability.

10.6.1 Finding the MAP solution

One obvious possibility is to find the maximum a posteriori (MAP) solution:

$$\begin{aligned}\hat{w}_{1\ldots N} &= \underset{w_{1\ldots N}}{\operatorname{argmax}}\left[Pr(w_{1\ldots N}|\mathbf{x}_{1\ldots N})\right] \\ &= \underset{w_{1\ldots N}}{\operatorname{argmax}}\left[Pr(\mathbf{x}_{1\ldots N}|w_{1\ldots N})Pr(w_{1\ldots N})\right].\end{aligned}$$

This is still far from trivial. The number of world states is extremely large so we cannot possibly explore every one and take the maximum. We must employ intelligent and efficient algorithms that exploit the redundancies in the distribution to find the correct solution where possible. However, as we shall see, for some models there is no known polynomial algorithm to find the MAP solution.

10.6.2 Finding the marginal posterior distribution

An alternative strategy is to find the marginal posterior distributions:

$$Pr(w_n|\mathbf{x}_{1\ldots N}) = \int\int Pr(w_{1\ldots N}|\mathbf{x}_{1\ldots N}) dw_{1\ldots n-1}\, dw_{n+1\ldots N}. \tag{10.15}$$

Since each of these distributions is over a single label, it is not implausible to compute and store each one separately. Obviously it is not possible to do this by directly computing the (extremely large) joint distribution and marginalizing it directly. We must use algorithms that exploit the conditional independence relations in the distribution to efficiently compute the marginals.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

10.6.3 Maximum marginals

If we want a single estimate of the world state, we could return the maximum values of the marginal distributions, giving the criterion

$$\hat{w}_n = \underset{w_n}{\operatorname{argmax}} \left[Pr(w_n|\mathbf{x}_{1\ldots N}) \right]. \tag{10.16}$$

This produces estimates of each world state that are individually most probable, but which may not reflect the joint statistics. For example, world state $w_n = 4$ might be the most probable value for the n^{th} world state and $w_m = 6$ might be the most probable value for the m^{th} world state, but it could be that the posterior probability for this configuration is zero: although the states are individually probable, they never co-occur (figure 10.10).

10.6.4 Sampling the posterior

For some models, it is intractable to compute either the MAP solution or the marginal distributions. One possibility in this circumstance is to draw samples from the posterior distribution. Methods based on sampling the posterior would fall under the more general category of *approximate inference* as they do not normally return the true answer.

Having drawn a number of samples from the posterior, we can approximate the posterior probability distribution as a mixture of delta functions where there is one delta function at each of the sample positions. Alternatively, we could make estimates of marginal statistics such as the mean and variance based on the sampled values or select the sample with the highest posterior probability as an estimate of the MAP state; this latter approach has the advantage of being consistent with the full posterior distribution (as opposed to maximum marginals which is not) even if we cannot be sure that we have the correct answer.

An alternative way to compute a point estimate from a set of samples from the posterior is to compute the *empirical max-marginals*. We estimate the marginal probability distributions by looking at the marginal statistics of the samples. In other words, we consider one variable w_n at a time and look at the distribution of different values observed. For a discrete distribution, this information is captured in a histogram. For a continuous distribution, we could fit a univariate model such as a normal distribution to these values to summarize them.

10.7 Drawing samples

We have seen that some of the approaches to inference require us to draw samples from the posterior distribution. We will now discuss how to do this for both directed and undirected models and we will see that this is generally more straightforward in directed models.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

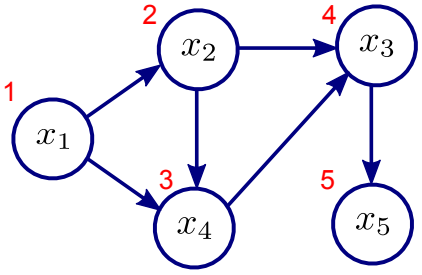

Figure 10.11 Ancestral sampling. We work our way through the graph in an order (red number) that guarantees that the parents of every node are visited before the node itself. At each step we draw a sample conditioned on the values of the samples at the parents. This is guaranteed to produce a valid sample from the full joint distribution.

10.7.1 Sampling from directed graphical models

Directed graphical models take the form of directed acyclic graphs of conditional probability relations that have the following algebraic form:

$$Pr(x_{1\ldots N}) = \prod_{n=1}^{I} Pr(x_n|x_{\mathrm{pa}[n]}). \tag{10.17}$$

It is relatively easy to sample from a directed graphical model using a technique known as *ancestral sampling*. The idea is to sample each variable in the network in turn, where the order is such that all parents of a node are sampled before the node itself. At each node, we condition on the observed sample values of the parents.

The simplest way to understand this is with an example. Consider the directed graphical model in figure 10.11 whose probability distribution factorizes as

$$Pr(x_1, x_2, x_3, x_4, x_5) = Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_4, x_2)Pr(x_4|x_2, x_1)Pr(x_5|x_3). \tag{10.18}$$

To sample from this model we first identify x_1 as a node with no parents and draw a sample from the distribution $Pr(x_1)$. Let us say the observed sample at x_1 took value α_1.

We now turn to the remaining nodes. Node x_2 is the only node in the network where all of the parents have been processed, and so we turn our attention here next. We draw a sample from the distribution $Pr(x_2|x_1 = \alpha_1)$ to yield a sample α_2. We now see that we are not yet ready to sample from x_3 as not all of its parents have been sampled, but we can sample x_4 from the distribution $Pr(x_4|x_1 = \alpha_1, x_2 = \alpha_2)$ to yield the value α_4. Continuing this process we draw x_3 from $Pr(x_3|x_2 = \alpha_2, x_4 = \alpha_4)$ and finally x_5 from $Pr(x_5|x_3 = \alpha_3)$.

The resulting vector $\mathbf{w}^* = [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5]$ is guaranteed to be a valid sample from the full joint distribution $Pr(x_1, x_2, x_3, x_4, x_5)$. An equivalent way to think about this algorithm is to consider it as working through the terms in the factorized joint distribution (right hand side of equation 10.18) sampling from each in turn conditioned on the previous values.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

10.7.2 Sampling from undirected graphical models

Unfortunately, it is much harder to draw samples from undirected models except in certain special cases (e.g., where the variables are continuous and Gaussian or where the graph structure takes the form of a tree). In general graphs we cannot use ancestral sampling because (i) there is no sense in which any variable is a parent to any other so we don't know which order to sample in and (ii) the terms $\phi[\bullet]$ in the factorization are not probability distributions anyway.

Algorithm 10.1

One way to generate samples from any complex high-dimensional probability distribution is to use a *Markov chain Monte Carlo* (MCMC) method. The principle is to generate a series (chain) of samples from the distribution, so that each sample depends directly on the previous one (hence "Markov"). However, the generation of the sample is not completely deterministic (hence "Monte Carlo").

One of the simplest MCMC methods is *Gibbs sampling* which proceeds as follows. First, we randomly choose the initial state $\mathbf{x}^{[0]}$ using any method. We generate the next sample in the chain $\mathbf{x}^{[1]}$ by updating the state at each dimension $\{x_n\}_{n=1}^{N}$ in turn (in any order). To update the n^{th} dimension x_n we fix the other $N-1$ dimensions and draw from the conditional distribution $Pr(x_n|x_{1\ldots N\backslash n})$ where the set $x_{1\ldots N\backslash n}$ denotes all of the N variables $x_1, x_2 \ldots x_N$ *except* x_n. Having modified every dimension in this way, we obtain the second sample in the chain. This idea is illustrated in figure 10.12 for the multivariate normal distribution.

When this procedure is repeated a very large number of times, so that the initial conditions are forgotten, a sample from this sequence can be considered as a draw from the distribution $Pr(x_{1\ldots N})$. Although this is not immediately obvious (and a proof is beyond the scope of this book) this procedure does clearly have some sensible properties: since we are sampling from the conditional probability distribution at each pixel, we are more likely to change the current value to one which has an overall higher probability. However, the stochastic update rule provides the possibility of (infrequently) visiting less probable regions of the space.

For undirected graphical models, the conditional distribution $Pr(x_n|x_{1\ldots N\backslash n})$ can be quite efficient to evaluate because of the conditional independence properties: variable x_n is conditionally independent of the rest of the nodes given its immediate neighbors, and so computing this term only involves the immediate neighbors. However, overall this method is extremely inefficient: it requires a large amount of computational effort to generate even a single sample. Sampling from directed graphical models is far easier.

10.8 Learning

In the section 10.7 we argued that sampling from directed graphical models is considerably easier than sampling from undirected graphical models. In this section, we consider learning in each type of model and come to a similar conclusion. Note that we are not discussing the learning of the graph structure here; we are talking about learning the parameters of the model itself. For directed graphical models, these parameters would determine the conditional distributions $Pr(x_n|x_{\mathrm{pa}[n]})$

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

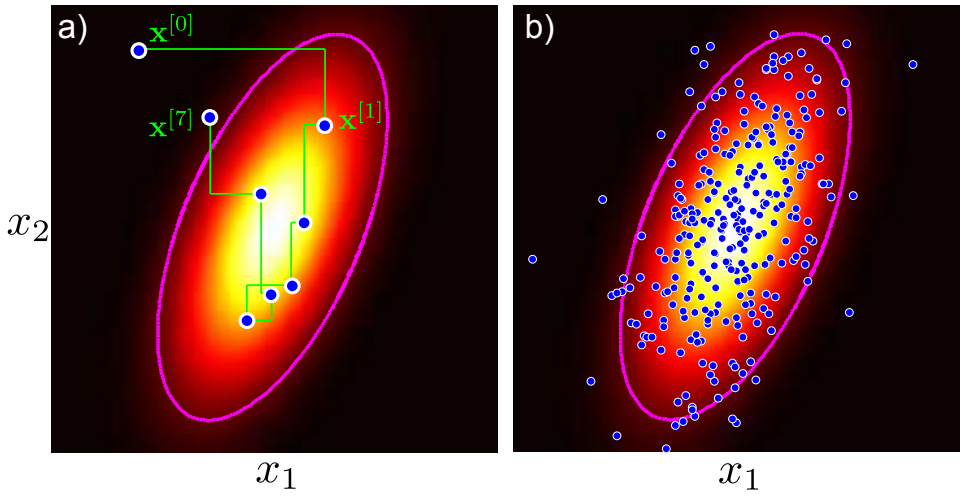

Figure 10.12 Gibbs sampling. We generate a chain of samples by cycling through each dimension in turn and drawing a sample from the conditional distribution of that dimension given that the others are fixed. a) For this 2D multivariate normal distribution we start at a random position $\mathbf{x}^{[0]}$. We alternately draw samples from the conditional distribution of the first dimension keeping the second fixed (horizontal changes) and the second dimension keeping the first fixed (vertical changes). For the multivariate normal, these conditional distributions are themselves normal (section 5.5). Each time we cycle through both of the dimensions, we create a new sample $\mathbf{x}^{[t]}$. b) Many samples generated using this method.

and for undirected graphical models they would determine the potential functions $\phi_c[\mathbf{x}_{1\ldots N}]$.

10.8.1 Learning in directed graphical models

Any directed graphical model can be written in the factorized form

$$Pr(x_1 \ldots x_N) = \prod_{n=1}^{N} Pr(x_n | x_{\mathrm{pa}[n]}, \boldsymbol{\theta}), \tag{10.19}$$

where the conditional probability relations form a directed acyclic graph, and $\boldsymbol{\theta}$ denotes the parameters of the model. For example, in the discrete distributions that we have focused on in this chapter, an individual conditional model might be

$$Pr(x_2 | x_1 = k) = \mathrm{Cat}_{x_2}[\boldsymbol{\lambda}_k] \tag{10.20}$$

where the parameters here are $\{\boldsymbol{\lambda}_k\}_{k=1}^{K}$. In general, the parameters can be learned using the maximum likelihood method by finding

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\prod_{i=1}^{I}\prod_{n=1}^{N} Pr(x_{i,n}|x_{i,\mathrm{pa}[n]},\boldsymbol{\theta})\right] \tag{10.21}$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\sum_{i=1}^{I}\sum_{n=1}^{N} \log[Pr(x_{i,n}|x_{i,\mathrm{pa}[n]},\boldsymbol{\theta})]\right], \tag{10.22}$$

where $x_{i,n}$ represents the n^{th} dimension of the i^{th} training example. This criterion leads to simple learning algorithms and often the maximum likelihood parameters can be computed in closed form.

10.8.2 Learning in undirected graphical models

An undirected graphical model is written as

$$Pr(\mathbf{x}) = \frac{1}{Z}\prod_{c=1}^{C}\phi_c[\mathbf{x},\boldsymbol{\theta}], \tag{10.23}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ and we have assumed that the training samples are independent. However, in this form we must constrain the parameters so that they ensure that each $\phi_c[\bullet]$ always returns a positive number. A more practical approach is to re-parameterize the undirected graphical model in the form of the Gibbs distribution,

$$Pr(\mathbf{x}) = \frac{1}{Z}\exp\left[-\sum_{c=1}^{C}\psi_c[x_{1\ldots N},\boldsymbol{\theta}]\right] \tag{10.24}$$

so that we do not have to worry about constraints on the parameters.

Given I training examples $\{\mathbf{x}_i\}_{i=1}^{I}$, we aim to fit parameters $\boldsymbol{\theta}$. Assuming that the training examples are independent, the maximum likelihood solution is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\frac{1}{Z(\boldsymbol{\theta})^I}\exp\left[-\sum_{i=1}^{I}\sum_{c=1}^{C}\psi_c(\mathbf{x}_i,\boldsymbol{\theta})\right]\right]$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[-I\log[Z(\boldsymbol{\theta})] - \sum_{i=1}^{I}\sum_{c=1}^{C}\psi_c(\mathbf{x}_i,\boldsymbol{\theta})\right], \tag{10.25}$$

where as usual we have taken the log to simplify the expression.

To maximize this expression we calculate the derivative of the log likelihood L with respect to the parameters $\boldsymbol{\theta}$:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -I\frac{\partial \log[Z(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} - \sum_{i=1}^{I}\sum_{c=1}^{C}\frac{\partial \psi_c(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{10.26}$$

$$= -I\frac{\partial \log\left[\sum_{\mathbf{x}_i}\exp\left[-\sum_{c=1}^{C}\psi_c(\mathbf{x}_i,\boldsymbol{\theta})\right]\right]}{\partial \boldsymbol{\theta}} - \sum_{i=1}^{I}\sum_{c=1}^{C}\frac{\partial \psi_c(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

The second term is readily computable but the first term involves an intractable sum over all possible values of the variable $\mathbf{x}$: we cannot compute the derivative with respect to the parameters for reasonable-sized models and so learning is difficult. Moreover, we cannot evaluate the original probability expression (equation 10.23) as this too contains an intractable sum. Consequently, we can't compute the derivative using finite differences either.

In short, we can neither find an algebraic solution nor use a straightforward optimization technique as we cannot compute the gradient. The best that we can do is to approximate the gradient.

Contrastive divergence

Algorithm 10.2

One possible solution to this problem is the *contrastive divergence* algorithm. This is a method for approximating the gradient of the log likelihood with respect to parameters $\boldsymbol{\theta}$ for functions with the general form,

$$Pr(\mathbf{x}) = \frac{1}{Z(\boldsymbol{\theta})} f[\mathbf{x}, \boldsymbol{\theta}], \tag{10.27}$$

where $Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}} f[\mathbf{x}, \boldsymbol{\theta}]$ is the normalizing constant and the derivative of the log likelihood is

$$\frac{\partial \log[Pr(\mathbf{x})]}{\partial \boldsymbol{\theta}} = -\frac{\partial \log[Z(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} + \frac{\partial \log[f[\mathbf{x}, \boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}}. \tag{10.28}$$

The main idea behind contrastive divergence follows from some algebraic manipulation of the first term:

$$\begin{aligned}
\frac{\partial \log[Z(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} &= \frac{1}{Z(\boldsymbol{\theta})} \frac{\partial Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{Z(\boldsymbol{\theta})} \frac{\partial \sum_{\mathbf{x}} f[\mathbf{x}, \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{Z(\boldsymbol{\theta})} \sum_{\mathbf{x}} \frac{\partial f[\mathbf{x}, \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{Z(\boldsymbol{\theta})} \sum_{\mathbf{x}} f[\mathbf{x}, \boldsymbol{\theta}] \frac{\partial \log[f[\mathbf{x}, \boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}} \\
&= \sum_{\mathbf{x}} Pr(\mathbf{x}) \frac{\partial \log[f[\mathbf{x}, \boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}}.
\end{aligned} \tag{10.29}$$

where we have used the relation $\partial \log f[\mathbf{x}]/\partial x = (\partial f[\mathbf{x}]/\partial x)/f[\mathbf{x}]$ between the third and fourth lines.

The final term in equation 10.29 is the expectation of the derivative of $\log[f[\mathbf{x}, \boldsymbol{\theta}]]$. We cannot compute this exactly, but we can approximate it by drawing J independent samples $\mathbf{x}^*$ from the distribution to yield

$$\frac{\partial \log[Z(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} = \sum_{\mathbf{x}} Pr(\mathbf{x}) \frac{\partial \log[f[\mathbf{x}, \boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}} \approx \frac{1}{J} \sum_{j=1}^{J} \frac{\partial \log[f[\mathbf{x}_j^*, \boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}}. \tag{10.30}$$

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

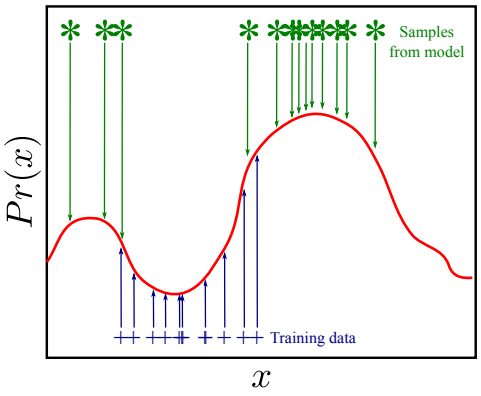

Figure 10.13 The contrastive divergence algorithm changes the parameters so that the un-normalized distribution increases at the observed data points (blue crosses) but decreases at sampled data points from the model. These two components counterbalance one another and ensure that the likelihood increases. When the model fits the data these two forces will cancel out and the parameters will remain constant.

With I training examples $\{\mathbf{x}_i\}_{i=1}^{I}$, the gradient of the log likelihood L is hence

$$\frac{\partial L}{\partial \boldsymbol{\theta}} \approx -\frac{I}{J}\sum_{j=1}^{J}\frac{\partial \log[f(\mathbf{x}_j^*, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} + \sum_{i=1}^{I}\frac{\partial \log[f(\mathbf{x}_i, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}. \tag{10.31}$$

A visual explanation of this expression is presented in figure 10.13. The gradient points in a direction that (i) increases the logarithm of the un-normalized function at the data points $\mathbf{x}_i$ but (ii) decreases the same quantity in places where the model believes the density is high (i.e., the samples $\mathbf{x}_j^*$). When the model fits the data, these two forces cancel out, and the parameters will stop changing.

This algorithm requires us to draw samples $\mathbf{x}^*$ from the model at each iteration of the optimization procedure in order to compute the gradient. Unfortunately, the only way to draw samples from general undirected graphical models is to use costly Markov chain Monte Carlo methods such as Gibbs sampling (section 10.7.2), and this is impractically time consuming. In practice it has been found that even approximate samples will do: one method is to re-start $J = I$ samples at the data points at each iteration and do just a few MCMC steps. Surprisingly, this works well even with a single step. A second approach is to start with the samples from the previous iteration and perform a few MCMC steps so that the samples are free to wander without restarting. This technique is known as *persistent* contrastive divergence.

Discussion

In this chapter, we introduced directed and undirected graphical models. Each represents a different type of factorization of the joint distribution. A graphical model implies a set of independence and conditional independence relations. There are some sets that can only be represented by directed graphical models, others that can only be represented by undirected graphical models, some that can be represented by both and some that cannot be represented by either.

We presented a number of common vision models and examined their graphical

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

models. Each had sparse connections and hence many conditional independence relations. In subsequent chapters, we will exploit these redundancies to develop efficient learning and inference algorithms. The world state is usually very high dimensional in these models and so we discussed alternative forms of inference including maximum marginals and sampling.

Finally, we looked at the implications of choosing directed or undirected graphical models for sampling and for learning. We concluded that it is generally more straightforward to draw samples from directed graphical models. Moreover, it is also easier to learn directed graphical models. The best-known learning algorithm for general undirected graphical models requires us to draw samples, which is itself challenging.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

Notes

Graphical models: For a readable introduction to graphical models, consult Jordan (2004) or Bishop (2006). For a more comprehensive overview, I would recommend Barber (2012). For an even more encyclopaedic resource, consult Koller & Friedman (2009).

Contrastive divergence: The contrastive divergence algorithm was introduced by Hinton (2002). Further information about this technique can be found in Carreira-Perpiñán. & Hinton (2005) and Bengio & Delalleau (2009).

Problems

Problem 10.1 The joint probability model between variables $\{x_n\}_{n=1}^{7}$ factorizes as

$$Pr(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = Pr(x_1)Pr(x_3)Pr(x_7)Pr(x_2|x_1, x_3)Pr(x_5|x_7, x_2)Pr(x_4|x_2)Pr(x_6|x_5, x_4).$$

Draw a directed graphical model relating these variables. Which variables form the Markov blanket of variable x_2?

Problem 10.2 Write out the factorization corresponding to the directed graphical model in figure 10.14a.

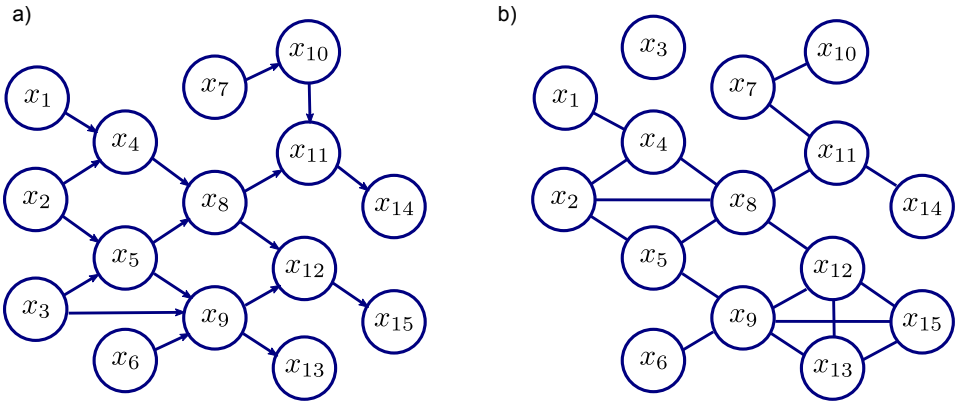

Figure 10.14 a) Graphical model for problem 10.2. b) Graphical model for problem 10.4

Problem 10.3 An undirected graphical model has the form

$$Pr(x_1 \ldots x_6) = \frac{1}{Z}\Phi_1[x_1, x_2, x_5]\Phi_2[x_2, x_3, x_4]\Phi_3[x_1x_5]\Phi_4[x_5, x_6].$$

Draw the undirected graphical model that corresponds to this factorization.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

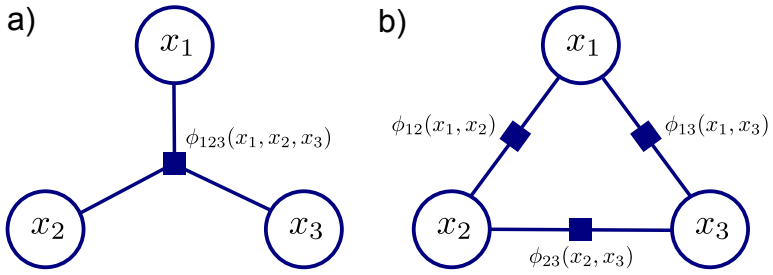

Figure 10.15 Factor graphs contain one node (square) per factor in the joint pdf as well as one node (circle) per variable. Each factor node is connected to all of the variables that belong to that factor. This type of graphical model can distinguish between the undirected graphical models a) $Pr(x_1, x_2, x_3) = \frac{1}{Z}\phi_{123}[x_1, x_2, x_3]$ and b) $Pr(x_1, x_2, x_3) = \frac{1}{Z}\phi_{12}[x_1, x_2]\phi_{23}[x_2, x_3]\phi_{13}[x_1, x_3]$.

Problem 10.4 Write out the factorization corresponding to the undirected graphical model in figure 10.14b.

Problem 10.5 Consider the undirected graphical model defined over binary values $\{x_i\}_{i=1}^{4} \in \{0, 1\}$ defined by

$$Pr(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi(x_1, x_2)\phi(x_2, x_3)\phi(x_3, x_4)\phi(x_4, x_1),$$

where the function ϕ is defined by

$$\begin{aligned} \phi(0,0) &= 1 & \phi(1,1) &= 2 \\ \phi(0,1) &= 0.1 & \phi(1,0) &= 0.1 \end{aligned}$$

Compute the probability of each of the 16 possible states of this system.

Problem 10.6 What is the Markov blanket for each of the variables in figures 10.7 and 10.8?

Problem 10.7 Show that the stated patterns of independence and conditional independence in figure 10.7 and figure 10.8 are true.

Problem 10.8 A *factor graph* is a third type of graphical model that depicts the factorization of a joint probability. As usual it contains a single node per variable, but it also contains one node per factor (usually indicated by a solid square). Each factor variable is connected to all of the variables that are contained in the associated term in the factorization by undirected links. For example, the factor node corresponding to the term $Pr(x_1|x_2, x_3)$ in a directed model would connect to all three variables x_1, x_2 and x_3. Similarly, the factor node corresponding to the term $\phi_{12}[x_1, x_2]$ in an undirected model would connect variables x_1 and x_2. Figure 10.15 shows two examples of factor graphs.

Draw the factor graphs corresponding to the graphical models in figures 10.7 and 10.8. You must first establish the factorized joint distribution associated with each graph.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

Problem 10.9 What is the Markov blanket of variable w_2 in figure 10.9c?

Problem 10.10 What is the Markov blanket of variable w_8 in figure 10.9e?

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.

Copyright ©2011,2012 by Simon Prince; published by Cambridge University Press 2012.
For personal use only, not for distribution.