

The image segmentation problem discussed above also provides an illustration of the use of clustering for data compression. Suppose the original image has  $N$  pixels comprising  $\{R, G, B\}$  values each of which is stored with 8 bits of precision. Then to transmit the whole image directly would cost  $24N$  bits. Now suppose we first run  $K$ -means on the image data, and then instead of transmitting the original pixel intensity vectors we transmit the identity of the nearest vector  $\mu_k$ . Because there are  $K$  such vectors, this requires  $\log_2 K$  bits per pixel. We must also transmit the  $K$  code book vectors  $\mu_k$ , which requires  $24K$  bits, and so the total number of bits required to transmit the image is  $24K + N \log_2 K$  (rounding up to the nearest integer). The original image shown in Figure 9.3 has  $240 \times 180 = 43,200$  pixels and so requires  $24 \times 43,200 = 1,036,800$  bits to transmit directly. By comparison, the compressed images require 43,248 bits ( $K = 2$ ), 86,472 bits ( $K = 3$ ), and 173,040 bits ( $K = 10$ ), respectively, to transmit. These represent compression ratios compared to the original image of 4.2%, 8.3%, and 16.7%, respectively. We see that there is a trade-off between degree of compression and image quality. Note that our aim in this example is to illustrate the  $K$ -means algorithm. If we had been aiming to produce a good image compressor, then it would be more fruitful to consider small blocks of adjacent pixels, for instance  $5 \times 5$ , and thereby exploit the correlations that exist in natural images between nearby pixels.

## 9.2. Mixtures of Gaussians

In Section 2.3.9 we motivated the Gaussian mixture model as a simple linear superposition of Gaussian components, aimed at providing a richer class of density models than the single Gaussian. We now turn to a formulation of Gaussian mixtures in terms of discrete *latent* variables. This will provide us with a deeper insight into this important distribution, and will also serve to motivate the expectation-maximization algorithm.

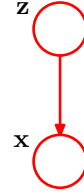
Recall from (2.188) that the Gaussian mixture distribution can be written as a linear superposition of Gaussians in the form

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k). \quad (9.7)$$

Let us introduce a  $K$ -dimensional binary random variable  $\mathbf{z}$  having a 1-of- $K$  representation in which a particular element  $z_k$  is equal to 1 and all other elements are equal to 0. The values of  $z_k$  therefore satisfy  $z_k \in \{0, 1\}$  and  $\sum_k z_k = 1$ , and we see that there are  $K$  possible states for the vector  $\mathbf{z}$  according to which element is nonzero. We shall define the joint distribution  $p(\mathbf{x}, \mathbf{z})$  in terms of a marginal distribution  $p(\mathbf{z})$  and a conditional distribution  $p(\mathbf{x} | \mathbf{z})$ , corresponding to the graphical model in Figure 9.4. The marginal distribution over  $\mathbf{z}$  is specified in terms of the mixing coefficients  $\pi_k$ , such that

$$p(z_k = 1) = \pi_k$$

**Figure 9.4** Graphical representation of a mixture model, in which the joint distribution is expressed in the form  $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ .



where the parameters  $\{\pi_k\}$  must satisfy

$$0 \leq \pi_k \leq 1 \quad (9.8)$$

together with

$$\sum_{k=1}^K \pi_k = 1 \quad (9.9)$$

in order to be valid probabilities. Because  $\mathbf{z}$  uses a 1-of- $K$  representation, we can also write this distribution in the form

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}. \quad (9.10)$$

Similarly, the conditional distribution of  $\mathbf{x}$  given a particular value for  $\mathbf{z}$  is a Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

which can also be written in the form

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}. \quad (9.11)$$

### Exercise 9.3

The joint distribution is given by  $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ , and the marginal distribution of  $\mathbf{x}$  is then obtained by summing the joint distribution over all possible states of  $\mathbf{z}$  to give

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (9.12)$$

where we have made use of (9.10) and (9.11). Thus the marginal distribution of  $\mathbf{x}$  is a Gaussian mixture of the form (9.7). If we have several observations  $\mathbf{x}_1, \dots, \mathbf{x}_N$ , then, because we have represented the marginal distribution in the form  $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$ , it follows that for every observed data point  $\mathbf{x}_n$  there is a corresponding latent variable  $\mathbf{z}_n$ .

We have therefore found an equivalent formulation of the Gaussian mixture involving an explicit latent variable. It might seem that we have not gained much by doing so. However, we are now able to work with the joint distribution  $p(\mathbf{x}, \mathbf{z})$

instead of the marginal distribution  $p(\mathbf{x})$ , and this will lead to significant simplifications, most notably through the introduction of the expectation-maximization (EM) algorithm.

Another quantity that will play an important role is the conditional probability of  $\mathbf{z}$  given  $\mathbf{x}$ . We shall use  $\gamma(z_k)$  to denote  $p(z_k = 1|\mathbf{x})$ , whose value can be found using Bayes' theorem

$$\begin{aligned}\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^K p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.\end{aligned}\quad (9.13)$$

We shall view  $\pi_k$  as the prior probability of  $z_k = 1$ , and the quantity  $\gamma(z_k)$  as the corresponding posterior probability once we have observed  $\mathbf{x}$ . As we shall see later,  $\gamma(z_k)$  can also be viewed as the *responsibility* that component  $k$  takes for ‘explaining’ the observation  $\mathbf{x}$ .

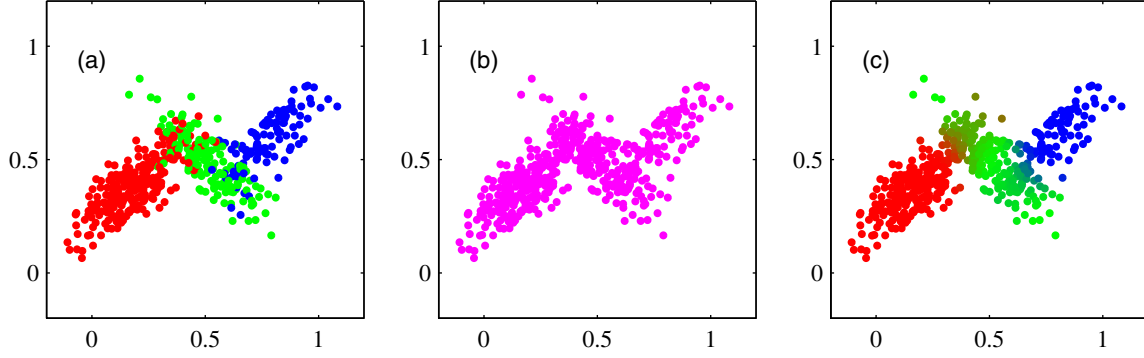
### Section 8.1.2

We can use the technique of ancestral sampling to generate random samples distributed according to the Gaussian mixture model. To do this, we first generate a value for  $\mathbf{z}$ , which we denote  $\hat{\mathbf{z}}$ , from the marginal distribution  $p(\mathbf{z})$  and then generate a value for  $\mathbf{x}$  from the conditional distribution  $p(\mathbf{x}|\hat{\mathbf{z}})$ . Techniques for sampling from standard distributions are discussed in Chapter 11. We can depict samples from the joint distribution  $p(\mathbf{x}, \mathbf{z})$  by plotting points at the corresponding values of  $\mathbf{x}$  and then colouring them according to the value of  $\mathbf{z}$ , in other words according to which Gaussian component was responsible for generating them, as shown in Figure 9.5(a). Similarly samples from the marginal distribution  $p(\mathbf{x})$  are obtained by taking the samples from the joint distribution and ignoring the values of  $\mathbf{z}$ . These are illustrated in Figure 9.5(b) by plotting the  $\mathbf{x}$  values without any coloured labels.

We can also use this synthetic data set to illustrate the ‘responsibilities’ by evaluating, for every data point, the posterior probability for each component in the mixture distribution from which this data set was generated. In particular, we can represent the value of the responsibilities  $\gamma(z_{nk})$  associated with data point  $\mathbf{x}_n$  by plotting the corresponding point using proportions of red, blue, and green ink given by  $\gamma(z_{nk})$  for  $k = 1, 2, 3$ , respectively, as shown in Figure 9.5(c). So, for instance, a data point for which  $\gamma(z_{n1}) = 1$  will be coloured red, whereas one for which  $\gamma(z_{n2}) = \gamma(z_{n3}) = 0.5$  will be coloured with equal proportions of blue and green ink and so will appear cyan. This should be compared with Figure 9.5(a) in which the data points were labelled using the true identity of the component from which they were generated.

#### 9.2.1 Maximum likelihood

Suppose we have a data set of observations  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , and we wish to model this data using a mixture of Gaussians. We can represent this data set as an  $N \times D$



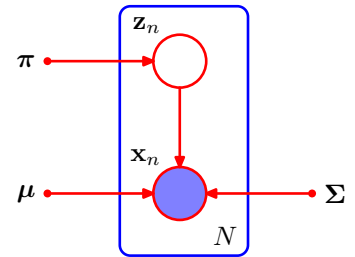
**Figure 9.5** Example of 500 points drawn from the mixture of 3 Gaussians shown in Figure 2.23. (a) Samples from the joint distribution  $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$  in which the three states of  $\mathbf{z}$ , corresponding to the three components of the mixture, are depicted in red, green, and blue, and (b) the corresponding samples from the marginal distribution  $p(\mathbf{x})$ , which is obtained by simply ignoring the values of  $\mathbf{z}$  and just plotting the  $\mathbf{x}$  values. The data set in (a) is said to be *complete*, whereas that in (b) is *incomplete*. (c) The same samples in which the colours represent the value of the responsibilities  $\gamma(z_{nk})$  associated with data point  $\mathbf{x}_n$ , obtained by plotting the corresponding point using proportions of red, blue, and green ink given by  $\gamma(z_{nk})$  for  $k = 1, 2, 3$ , respectively

matrix  $\mathbf{X}$  in which the  $n^{\text{th}}$  row is given by  $\mathbf{x}_n^T$ . Similarly, the corresponding latent variables will be denoted by an  $N \times K$  matrix  $\mathbf{Z}$  with rows  $\mathbf{z}_n^T$ . If we assume that the data points are drawn independently from the distribution, then we can express the Gaussian mixture model for this i.i.d. data set using the graphical representation shown in Figure 9.6. From (9.7) the log of the likelihood function is given by

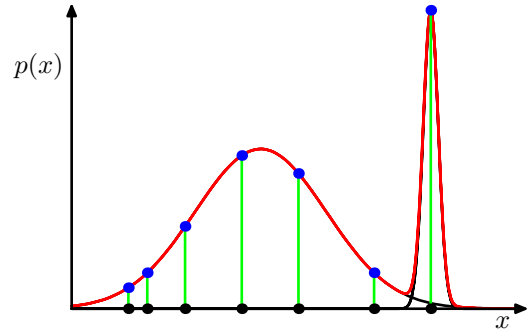
$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}. \quad (9.14)$$

Before discussing how to maximize this function, it is worth emphasizing that there is a significant problem associated with the maximum likelihood framework applied to Gaussian mixture models, due to the presence of singularities. For simplicity, consider a Gaussian mixture whose components have covariance matrices given by  $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}$ , where  $\mathbf{I}$  is the unit matrix, although the conclusions will hold for general covariance matrices. Suppose that one of the components of the mixture model, let us say the  $j^{\text{th}}$  component, has its mean  $\boldsymbol{\mu}_j$  exactly equal to one of the data

**Figure 9.6** Graphical representation of a Gaussian mixture model for a set of  $N$  i.i.d. data points  $\{\mathbf{x}_n\}$ , with corresponding latent points  $\{\mathbf{z}_n\}$ , where  $n = 1, \dots, N$ .



**Figure 9.7** Illustration of how singularities in the likelihood function arise with mixtures of Gaussians. This should be compared with the case of a single Gaussian shown in Figure 1.14 for which no singularities arise.



points so that  $\mu_j = \mathbf{x}_n$  for some value of  $n$ . This data point will then contribute a term in the likelihood function of the form

$$\mathcal{N}(\mathbf{x}_n | \mathbf{x}_n, \sigma_j^2 \mathbf{I}) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}. \quad (9.15)$$

If we consider the limit  $\sigma_j \rightarrow 0$ , then we see that this term goes to infinity and so the log likelihood function will also go to infinity. Thus the maximization of the log likelihood function is not a well posed problem because such singularities will always be present and will occur whenever one of the Gaussian components ‘collapses’ onto a specific data point. Recall that this problem did not arise in the case of a single Gaussian distribution. To understand the difference, note that if a single Gaussian collapses onto a data point it will contribute multiplicative factors to the likelihood function arising from the other data points and these factors will go to zero exponentially fast, giving an overall likelihood that goes to zero rather than infinity. However, once we have (at least) two components in the mixture, one of the components can have a finite variance and therefore assign finite probability to all of the data points while the other component can shrink onto one specific data point and thereby contribute an ever increasing additive value to the log likelihood. This is illustrated in Figure 9.7. These singularities provide another example of the severe over-fitting that can occur in a maximum likelihood approach. We shall see that this difficulty does not occur if we adopt a Bayesian approach. For the moment, however, we simply note that in applying maximum likelihood to Gaussian mixture models we must take steps to avoid finding such pathological solutions and instead seek local maxima of the likelihood function that are well behaved. We can hope to avoid the singularities by using suitable heuristics, for instance by detecting when a Gaussian component is collapsing and resetting its mean to a randomly chosen value while also resetting its covariance to some large value, and then continuing with the optimization.

### Section 10.1

A further issue in finding maximum likelihood solutions arises from the fact that for any given maximum likelihood solution, a  $K$ -component mixture will have a total of  $K!$  equivalent solutions corresponding to the  $K!$  ways of assigning  $K$  sets of parameters to  $K$  components. In other words, for any given (nondegenerate) point in the space of parameter values there will be a further  $K! - 1$  additional points all of which give rise to exactly the same distribution. This problem is known as

*identifiability* (Casella and Berger, 2002) and is an important issue when we wish to interpret the parameter values discovered by a model. Identifiability will also arise when we discuss models having continuous latent variables in Chapter 12. However, for the purposes of finding a good density model, it is irrelevant because any of the equivalent solutions is as good as any other.

Maximizing the log likelihood function (9.14) for a Gaussian mixture model turns out to be a more complex problem than for the case of a single Gaussian. The difficulty arises from the presence of the summation over  $k$  that appears inside the logarithm in (9.14), so that the logarithm function no longer acts directly on the Gaussian. If we set the derivatives of the log likelihood to zero, we will no longer obtain a closed form solution, as we shall see shortly.

One approach is to apply gradient-based optimization techniques (Fletcher, 1987; Nocedal and Wright, 1999; Bishop and Nabney, 2008). Although gradient-based techniques are feasible, and indeed will play an important role when we discuss mixture density networks in Chapter 5, we now consider an alternative approach known as the EM algorithm which has broad applicability and which will lay the foundations for a discussion of variational inference techniques in Chapter 10.

### 9.2.2 EM for Gaussian mixtures

An elegant and powerful method for finding maximum likelihood solutions for models with latent variables is called the *expectation-maximization* algorithm, or *EM* algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 1997). Later we shall give a general treatment of EM, and we shall also show how EM can be generalized to obtain the variational inference framework. Initially, we shall motivate the EM algorithm by giving a relatively informal treatment in the context of the Gaussian mixture model. We emphasize, however, that EM has broad applicability, and indeed it will be encountered in the context of a variety of different models in this book.

Let us begin by writing down the conditions that must be satisfied at a maximum of the likelihood function. Setting the derivatives of  $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$  in (9.14) with respect to the means  $\boldsymbol{\mu}_k$  of the Gaussian components to zero, we obtain

$$0 = - \sum_{n=1}^N \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\underbrace{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}_{\gamma(z_{nk})}} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \quad (9.16)$$

where we have made use of the form (2.43) for the Gaussian distribution. Note that the posterior probabilities, or responsibilities, given by (9.13) appear naturally on the right-hand side. Multiplying by  $\boldsymbol{\Sigma}_k^{-1}$  (which we assume to be nonsingular) and rearranging we obtain

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \quad (9.17)$$

where we have defined

$$N_k = \sum_{n=1}^N \gamma(z_{nk}). \quad (9.18)$$