representing the mixing parameters in terms of a set of M auxiliary variables $\{\gamma_j\}$ such that

$$P(j) = \frac{\exp(\gamma_j)}{\sum_{k=1}^{M} \exp(\gamma_k)}. \tag{2.81}$$

The transformation given by (2.81) is called the *softmax* function, or normalized exponential, and ensures that, for $-\infty \leq \gamma_j \leq \infty$, the constraints (2.72) and (2.73) are satisfied as required for probabilities. We can now perform an unconstrained minimization of the error function with respect to the $\{\gamma_j\}$. To find the derivatives of E with respect to γ_j we make use of

$$\frac{\partial P(k)}{\partial \gamma_j} = \delta_{jk} P(j) - P(j)P(k) \tag{2.82}$$

which follows from (2.81). Using the chain rule in the form

$$\frac{\partial E}{\partial \gamma_j} = \sum_{k=1}^{M} \frac{\partial E}{\partial P(k)} \frac{\partial P(k)}{\partial \gamma_j} \tag{2.83}$$

together with (2.75) and (2.78), we then obtain the required derivatives in the form

$$\frac{\partial E}{\partial \gamma_j} = -\sum_{n=1}^{N} \{P(j|\mathbf{x}^n) - P(j)\} \tag{2.84}$$

where we have made use of (2.76). The complete set of derivatives of the error function with respect to the parameters of the model, given by (2.79), (2.80) and (2.84), can then be used in the non-linear optimization algorithms described in Chapter 7 to provide practical techniques for finding minima of the error function.

Some insight into the nature of the maximum likelihood solution can be obtained by considering the expressions for the parameters at a minimum of E. Setting (2.79) to zero we obtain

$$\widehat{\boldsymbol{\mu}}_j = \frac{\sum_n P(j|\mathbf{x}^n)\mathbf{x}^n}{\sum_n P(j|\mathbf{x}^n)} \tag{2.85}$$

which represents the intuitively satisfying result that the mean of the jth component is just the mean of the data vectors, weighted by the posterior probabilities that the corresponding data points were generated from that component. Similarly, setting the derivatives in (2.80) to zero we find

$$\widehat{\sigma}_j^2 = \frac{1}{d} \frac{\sum_n P(j|\mathbf{x}^n)\|\mathbf{x}^n - \widehat{\boldsymbol{\mu}}_j\|^2}{\sum_n P(j|\mathbf{x}^n)} \tag{2.86}$$

which again represents the intuitive result that the variance of the jth component is given by the variance of the data with respect to the mean of that component, again weighted with the posterior probabilities. Finally, setting the derivative in (2.84) to zero we obtain

$$\widehat{P}(j) = \frac{1}{N} \sum_{n=1}^{N} P(j|\mathbf{x}^n) \tag{2.87}$$

so that, at the maximum likelihood solution, the prior probability for the jth component is given by the posterior probabilities for that component, averaged over the data set.

2.6.2 *The EM algorithm*

While the formulae given in (2.85), (2.86) and (2.87) provide useful insight into the nature of the maximum likelihood solution, they do not provide a direct method for calculating the parameters. In fact they represent highly non-linear coupled equations, since the parameters occur implicitly on the right-hand sides by virtue of (2.75). They do, however, suggest that we might seek an iterative scheme for finding the minima of E. Suppose we begin by making some initial guess for the parameters of the Gaussian mixture model, which we shall call the 'old' parameter values. We can then evaluate the right-hand sides in (2.85), (2.86) and (2.87), and this will give a revised estimate for the parameters, which we shall call the 'new' parameter values, for which we might hope the value of the error function is smaller. These parameter values then become the 'old' values, and the process is repeated. We shall show that, provided some care is taken over the way in which the updates are performed, an algorithm of this form can be found which is guaranteed to decrease the error function at each iteration, until a local minimum is found. This provides a simple, practical method for estimating the mixture parameters which avoids the complexities of non-linear optimization algorithms. We shall also see that this is a special case of a more general procedure known as the *expectation–maximization*, or EM, algorithm (Dempster *et al.*, 1977).

From (2.78) we can write the change in error when we replace the old parameter values by the new values in the form

$$E^{\text{new}} - E^{\text{old}} = -\sum_n \ln \left\{ \frac{p^{\text{new}}(\mathbf{x}^n)}{p^{\text{old}}(\mathbf{x}^n)} \right\} \tag{2.88}$$

where $p^{\text{new}}(\mathbf{x})$ denotes the probability density evaluated using the new values for the parameters, while $p^{\text{old}}(\mathbf{x})$ represents the density evaluated using the old parameter values. Using the definition of the mixture distribution given by (2.71),