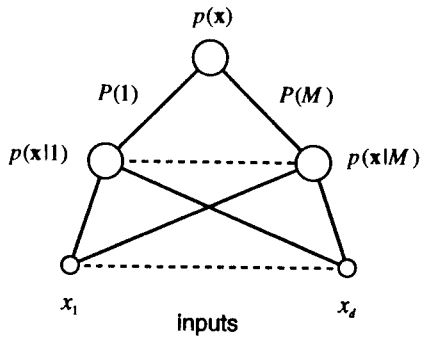

Figure 2.12. Representation of the mixture model (2.71) in terms of a network diagram. For Gaussian component densities $p(\mathbf{x}|j)$ given by (2.77), the lines connecting the inputs x_i to the components $p(\mathbf{x}|j)$ represent the elements μ_{ji} of the corresponding mean vectors $\boldsymbol{\mu}_j$.

The value of $P(j|\mathbf{x})$ represents the probability that a particular component j was responsible for generating the data point $\mathbf{x}$.

In this section, we shall limit our attention to mixture models in which the individual component densities are given by Gaussian distribution functions. We shall further assume that the Gaussians each have a covariance matrix which is some scalar multiple of the identity matrix so that $\boldsymbol{\Sigma}_j = \sigma_j^2\mathbf{I}$ (where $\mathbf{I}$ is the identity matrix) and hence

$$p(\mathbf{x}|j) = \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\left\{-\frac{\|\mathbf{x}-\boldsymbol{\mu}_j\|^2}{2\sigma_j^2}\right\}. \tag{2.77}$$

In fact, the techniques we shall describe are easily extended to general Gaussian component densities having full covariance matrices as discussed in Section 2.1.1 in the context of parametric distributions.

The mixture model can be represented in terms of a network diagram as shown in Figure 2.12. This is simply a diagrammatic representation of a mathematical function, in this case the mixture model in (2.71). Such diagrams prove particularly useful when considering complex neural network structures, as discussed in later chapters.

2.6.1 *Maximum likelihood*

Various procedures have been developed for determining the parameters of a Gaussian mixture model from a set of data. In the remainder of this chapter we consider three approaches, all of them based on maximizing the likelihood of the parameters for the given data set. A review of maximum likelihood techniques

in this context has been given by Redner and Walker (1984).

For the case of Gaussian components of the form (2.77), the mixture density contains the following adjustable parameters: $P(j)$, $\boldsymbol{\mu}_j$ and σ_j (where $j = 1,\ldots,M$). The negative log-likelihood for the data set is given by

$$E = -\ln\mathcal{L} = -\sum_{n=1}^{N}\ln p(\mathbf{x}^n) = -\sum_{n=1}^{N}\ln\left\{\sum_{j=1}^{M}p(\mathbf{x}^n|j)P(j)\right\} \tag{2.78}$$

which can be regarded as an error function. Maximizing the likelihood $\mathcal{L}$ is then equivalent to minimizing E.

It is important to emphasize that minimizing this error function is non-trivial in a number of respects. First of all, there exist parameter values for which the likelihood goes to infinity (Day, 1969). These arise when one of the Gaussian components collapses onto one of the data points, as can be seen by setting $\boldsymbol{\mu}_j = \mathbf{x}$ in (2.77) and then letting $\sigma_j \to 0$. In addition, small groups of points which are close together can give rise to local minima in the error function which may give poor representations of the true distribution. In practical problems we wish to avoid the singular solutions and the inappropriate local minima. Several techniques for dealing with the problems of singularities have been proposed. One approach is to constrain the components to have equal covariance matrices (Day, 1969). Alternatively, when one of the variance parameters shrinks to a small value during the course of an iterative algorithm, the corresponding Gaussian can be replaced with one having a larger width.

Since the error function is a smooth differentiable function of the parameters of the mixture model, we can employ standard non-linear optimization techniques, such as those described in Chapter 7, to find its minima. We shall see in Chapter 7, that there are considerable computational advantages in making use of gradient information provided it can be evaluated efficiently. In the present case the derivatives of E can be found analytically.

For the centres $\boldsymbol{\mu}_j$ of the Gaussian components we find, by simple differentiation of (2.78), and making use of (2.75) and (2.77),

$$\frac{\partial E}{\partial \boldsymbol{\mu}_j} = \sum_{n=1}^{N} P(j|\mathbf{x}^n)\frac{(\boldsymbol{\mu}_j - \mathbf{x}^n)}{\sigma_j^2}. \tag{2.79}$$

Similarly, for the width parameter σ_j we obtain

$$\frac{\partial E}{\partial \sigma_j} = \sum_{n=1}^{N} P(j|\mathbf{x}^n)\left\{\frac{d}{\sigma_j} - \frac{\|\mathbf{x}^n - \boldsymbol{\mu}_j\|^2}{\sigma_j^3}\right\}. \tag{2.80}$$

The minimization of E with respect to the mixing parameters $P(j)$ must be carried out subject to the constraints (2.72) and (2.73). This can be done by