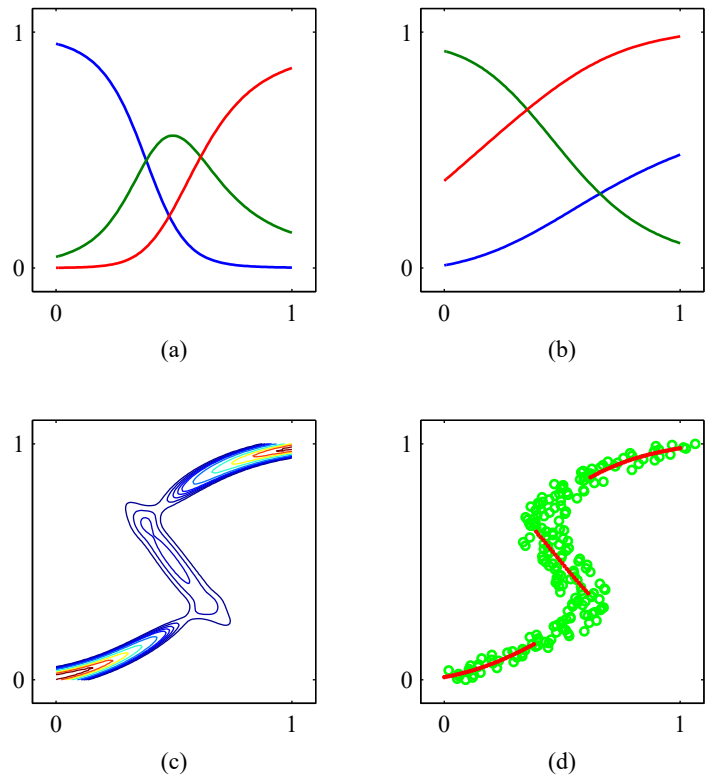

Figure 6.19 (a) Plot of the mixing coefficients $\pi_k(x)$ as a function of x for the three mixture components in a mixture density network trained on the data shown in Figure 6.17. The model has three Gaussian components and uses a two-layer neural network with five $\tanh$ sigmoidal units in the hidden layer and nine outputs (corresponding to the three means and three variances of the Gaussian components and the three mixing coefficients). At both small and large values of x, where the conditional probability density of the target data is unimodal, only one of the Gaussian components has a high value for its prior probability, whereas at intermediate values of x, where the conditional density is trimodal, the three mixing coefficients have comparable values. (b) Plots of the means $\mu_k(x)$ using the same colour coding as for the mixing coefficients. (c) Plot of the contours of the corresponding conditional probability density of the target data for the same mixture density network. (d) Plot of the approximate conditional mode, shown by the red points, of the conditional density.

where we have used (6.38). Because a standard network trained by least squares approximates the conditional mean, we see that a mixture density network can reproduce the conventional least-squares result as a special case. Of course, as we have already noted, for a multimodal distribution the conditional mean is of limited value.

Exercise 6.21

We can similarly evaluate the variance of the density function about the conditional average, to give

$$s^2(\mathbf{x}) = \mathbb{E}\left[\|\mathbf{t} - \mathbb{E}[\mathbf{t}|\mathbf{x}]\|^2 \,|\mathbf{x}\right] \tag{6.49}$$

$$= \sum_{k=1}^{K} \pi_k(\mathbf{x}) \left\{ \sigma_k^2(\mathbf{x}) + \left\| \boldsymbol{\mu}_k(\mathbf{x}) - \sum_{l=1}^{K} \pi_l(\mathbf{x}) \boldsymbol{\mu}_l(\mathbf{x}) \right\|^2 \right\} \tag{6.50}$$

where we have used (6.38) and (6.48). This is more general than the corresponding least-squares result because the variance is a function of $\mathbf{x}$.

We have seen that for multimodal distributions, the conditional mean can give a poor representation of the data. For instance, in controlling the simple robot arm shown in Figure 6.16, we need to pick one of the two possible joint angle settings