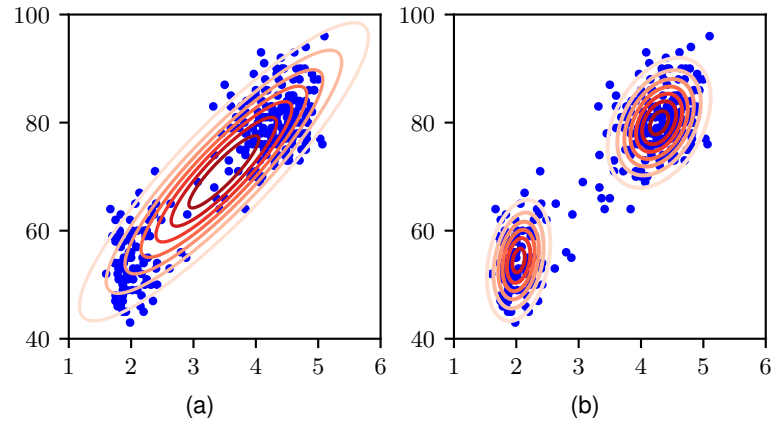

Figure 3.6 Plots of the Old Faithful data in which the red curves are contours of constant probability density. (a) A single Gaussian distribution which has been fitted to the data using maximum likelihood. Note that this distribution fails to capture the two clumps in the data and indeed places much of its probability mass in the central region between the clumps where the data are relatively sparse. (b) The distribution given by a linear combination of two Gaussians, also fitted by maximum likelihood, which gives a better representation of the data.

dissect out the contribution from the final data point $\mathbf{x}_N$, we obtain

$$\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\
&= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}).
\end{aligned} \tag{3.110}$$

This result has a nice interpretation, as follows. After observing $N-1$ data points, we estimate $\boldsymbol{\mu}$ by $\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}$. We now observe data point $\mathbf{x}_N$, and we obtain our revised estimate $\boldsymbol{\mu}_{\mathrm{ML}}^{(N)}$ by moving the old estimate a small amount, proportional to $1/N$, in the direction of the 'error signal' $(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)})$. Note that, as N increases, so the contributions from successive data points get smaller.

3.2.9 Mixtures of Gaussians

Although the Gaussian distribution has some important analytical properties, it suffers from significant limitations when used to model modelling real data sets. Consider the example shown in Figure 3.6(a). This is known as the 'Old Faithful' data set, and comprises 272 measurements of the eruption of the Old Faithful geyser in Yellowstone National Park in the USA. Each measurement gives the duration of the eruption in minutes (horizontal axis) and the time in minutes to the next eruption (vertical axis). We see that the data set forms two dominant clumps, and that a simple Gaussian distribution is unable to capture this structure.

We might expect that a superposition of two Gaussian distributions would be able to do a much better job of representing the structure in this data set, and indeed