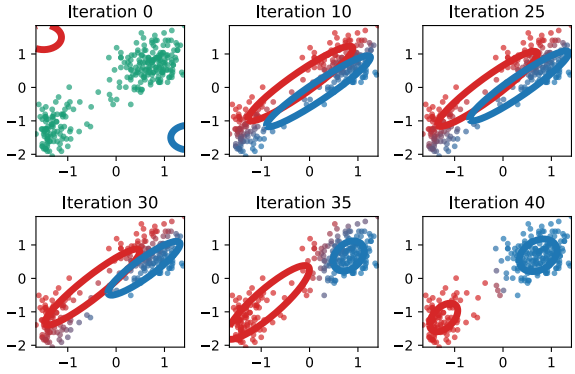

Figure 8.25: Illustration of the EM for a GMM applied to the Old Faithful data. The degree of redness indicates the degree to which the point belongs to the red cluster, and similarly for blue; thus purple points have a roughly 50/50 split in their responsibilities to the two clusters. Adapted from [Bis06] Figure 9.8. Generated by mix_gauss_demo_faithful.ipynb.

8.7.3.4 MAP estimation

Computing the MLE of a GMM often suffers from numerical problems and overfitting. To see why, suppose for simplicity that $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}$ for all k. It is possible to get an infinite likelihood by assigning one of the centers, say $\boldsymbol{\mu}_k$, to a single data point, say $\boldsymbol{y}_n$, since then the likelihood of that data point is given by

$$\mathcal{N}(\boldsymbol{y}_n | \boldsymbol{\mu}_k = \boldsymbol{y}_n, \sigma_k^2 \mathbf{I}) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^0 \tag{8.168}$$

Hence we can drive this term to infinity by letting $\sigma_k \to 0$, as shown in Figure 8.26(a). We call this the "collapsing variance problem".

An easy solution to this is to perform MAP estimation. Fortunately, we can still use EM to find this MAP estimate. Our goal is now to maximize the expected complete data log-likelihood plus the log prior:

$$\ell^t(\boldsymbol{\theta}) = \left[\sum_n \sum_k r_{nk}^{(t)} \log \pi_{nk} + \sum_n \sum_k r_{nk}^{(t)} \log p(\boldsymbol{y}_n | \boldsymbol{\theta}_k) \right] + \log p(\boldsymbol{\pi}) + \sum_k \log p(\boldsymbol{\theta}_k) \tag{8.169}$$

Note that the E step remains unchanged, but the M step needs to be modified, as we now explain.

For the prior on the mixture weights, it is natural to use a Dirichlet prior (Section 4.6.3.2), $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha})$, since this is conjugate to the categorical distribution. The MAP estimate is given by

$$\tilde{\pi}_k^{(t+1)} = \frac{r_k^{(t)} + \alpha_k - 1}{N + \sum_k \alpha_k - K} \tag{8.170}$$

If we use a uniform prior, $\alpha_k = 1$, this reduces to the MLE.