1 Introduction to EM

Consider a distribution $P(X|\Theta)$ parameterized by Θ, where X are data variables that are observed. Given data X, we want to find model parameters that agree with the data. In the maximum-likelihood framework, these are

$$\Theta^{ML} = \text{argmax}_{\Theta} P(X|\Theta)$$

If our model has latent variables Z that are not observed, and the joint distribution is $P(X, Z|\Theta)$, then the ML task is:

$$\Theta^{ML} = \text{argmax}_{\Theta} \sum_{Z} P(X, Z|\Theta)$$

In general this is difficult to optimize. One could try gradient ascent, but EM is an effective alternative.

If we had the Z, then maximizing the 'complete-likelihood' $P(X, Z|\Theta)$ would be easy. Therefore, we use a distribution over Z, denoted $Q(Z)$ and maximize the expected complete-likelihood with respect to this distribution:

$$E_{Z\sim Q(Z)}\left[P(X, Z|\Theta)\right]$$

If we had $Q(Z) = P(Z|X, \Theta)$, then this expectation is the same as our actual objective.

EM is an iterative algorithm. Throughout, we maintain a distribution over the hidden variables Z, denoted $q(Z)$.

2 EM for Gaussian Mixture Model

Using a one-of-k representation for the latent variables, the joint distribution is:

$$\begin{aligned} P(X, Z|\Theta) &= \prod_{n=1}^{N} P(x_n|z_n, \mu, \Sigma)P(z_n|\pi) \\ &= \prod_{n=1}^{N}\prod_{k=1}^{K} \left(N(x_n|\mu_k, \Sigma_k)\pi_k\right)^{z_{nk}} \end{aligned}$$

The complete log-likelihood is:

$$\log P(X, Z|\Theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left(\log \pi_k + \log N(x_n|\mu_k, \Sigma_k)\right)$$

Given a distribution over the Z represented by $\gamma_{nk} = P(z_{nk} = 1|X, \Theta^{old})$, the expected complete-log-likelihood is:

$$Q(\Theta, \Theta^{old}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left(\log \pi_k + \log N(x_n|\mu_k, \Sigma_k) \right)$$

Maximizing this with respect to Θ gives