More on the EM algorithm

Continue reading Chapter 9 in the text by Bishop
The EM Algorithm for Gaussian Mixture Models

Recall from last lecture the EM algorithm for a Gaussian mixture model with $\Sigma_k$ being diagonal, with diagonal elements of $\sigma_{kj}^2$.

The algorithm alternates between “E” steps and “M” steps:

**E Step:** Using the current values of the parameters, compute the “responsibilities” of components for data items, by applying Bayes’ Rule:

$$r_{ik} = P(\text{data item } i \text{ came from component } k \mid x_i) = \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} N(x_i \mid \mu_{k'}, \Sigma_{k'})}$$

**M Step:** Using the current responsibilities, re-estimate the parameters, using weighted averages, with weights given by the responsibilities:

$$\pi_k = \frac{1}{N} \sum_i r_{ik}, \quad \mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}, \quad \sigma_k^2 = \frac{\sum_i r_{ik} (x_i - \mu_k)^2}{\sum_i r_{ik}}$$

We start with some initial guess at the parameter values (perhaps random), or perhaps with some initial guess at the responsibilities (in which case we start with an M step). We continue alternating E and M steps until there is little change.
The EM Algorithm in General

Consider model for observed data $x$ (which might be a vector of $n$ independent items) that is accompanied by a latent (unobserved) $z$ (also possibly a vector of $n$ independent values). A model with parameters $\theta$ describes the joint distribution of $x$ and $z$, as $P(x, z|\theta)$.

We want to estimate $\theta$ by maximum likelihood, which means finding the $\theta$ that maximizes

$$P(x|\theta) = \sum_{z} P(x, z|\theta)$$

(This assumes $z$ is discrete; if it’s continuous the sum is replaced by an integral.) We assume that this isn’t easy. But suppose that we can easily find the $\theta$ that maximizes $P(x, z|\theta)$, for any known $x$ and $z$. We try to use (something related to) this capability in an iterative algorithm for maximizing $P(x|\theta)$. 
The EM Algorithm in General — Details

The general EM algorithm alternates these steps:

---

**E Step:** Using the current value of the parameter, $\theta$, find the distribution, $Q$, for the latent $z$, given the observed $x$:

$$Q(z) = P(z|x, \theta)$$

**M Step:** Maximize the expected value of $\log P(x, z|\theta)$ with respect to $\theta$, where the expectation is with respect to the distribution $Q$ found in the E step:

$$\theta = \arg \max_{\theta} E_Q[\log P(x, z|\theta)]$$

---

For many models (specifically, those in the “exponential family”), maximizing $E_Q[\log P(x, z|\theta)]$ will be feasible if maximizing $\log P(x, z|\theta)$ for known $z$ is feasible.
Justification of the EM algorithm

To see that the EM algorithm maximizes (at least locally) the log likelihood, consider the following function of the distribution $Q$ over $z$ and the parameters $\theta$:

$$F(Q, \theta) = E_Q[\log P(x, z|\theta)] - E_Q[\log Q(z)]$$

$$= \log P(x|\theta) + E_Q[\log P(z|x, \theta)] - E_Q[\log Q(z)]$$

$$= \log P(x|\theta) - E_Q[\log(Q(z)/P(z|x, \theta))]$$

The final term above is the “Kullback-Leibler (KL) divergence” between the distribution $Q(z)$ and the distribution $P(z|x, \theta)$. One can show that this divergence is always non-negative, and is zero only when $Q(z) = P(z|x, \theta)$.

We can now justify the EM algorithm by showing that

a) The E step maximizes $F(Q, \theta)$ with respect to $Q$ — a consequence of KL divergence being minimized when $Q(z) = P(z|x, \theta)$.

b) The M step maximizes $F(Q, \theta)$ with respect to $\theta$ — clear since $E_Q[\log Q(z)]$ doesn’t depend on $\theta$.

c) The maximum of $F(Q, \theta)$ occurs at a $\theta$ that maximizes $P(x|\theta)$ — if instead $P(x|\theta^*) > P(x|\theta)$ for some $\theta^*$, then $F(Q^*, \theta^*) > F(Q, \theta)$ with $Q^*(z) = P(z|x, \theta^*)$. 

How this Translates to the Mixture Version

For the mixture example, the model parameters are $\theta = (\pi, \mu, \sigma)$.

We’ll let the latent variables be $z_{ik} = 1$ if data item $i$ comes from component $k$, and 0 otherwise.

In the E step, we find the distribution of the $z_{ik}$ given $x_i$ and the model parameters. It turns out that all we actually need from this distribution is the expected value of each $z_{ik}$ (same as the probability that $z_{ik} = 1$), which we define to be $r_{ik}$, and find by Bayes’ Rule as shown before.

In the M step, we need to maximize $E_Q \left( \sum_{i=1}^{N} \log P(x_i, z_i | \theta) \right)$.

Suppose we knew the value of both $x_i$ and $z_i = (z_{i1}, \ldots, z_{iK})$ for data item $i$. The log probability (dropping constant factors) for that item can be written as

$$
\log \left[ \prod_{k=1}^{K} \left( \pi_k \prod_{j=1}^{D} \left( \frac{1}{\sigma_{kj}} \exp\left(-\frac{1}{2}(x_{ij} - \mu_{kj})^2 / \sigma_{kj}^2 \right) \right) \right)^{z_{ik}} \right]
$$

Note that all but one factor in the outer product will have the value one.

We maximize the expected value of the sum of the above for all $i$, with respect to the distribution of $z_i$ found in the E step. We’ll see how this works out next...
Details of the Mixture Version of EM

Taking the expectation of the log probability of data item $i$ with respect to the distribution of $z_i$ (denoted by $Q$), we get

$$E_Q \left\{ \log \left[ \prod_{k=1}^{K} \left( \prod_{j=1}^{D} \left( \frac{1}{\sigma_{k,j}} \exp\left(-\frac{1}{2}(x_{ij} - \mu_{k,j})^2 / \sigma_{k,j}^2 \right) \right) \right)^{z_{ik}} \right] \right\}$$

$$= E_Q \left\{ \sum_{k=1}^{K} z_{ik} \left( \log(\pi_k) - \frac{1}{2} \sum_{j=1}^{D} \left( \log(\sigma_{k,j}^2) + (x_{ij} - \mu_{k,j})^2 / \sigma_{k,j}^2 \right) \right) \right\}$$

$$= \sum_{k=1}^{K} r_{ik} \left( \log(\pi_k) - \frac{1}{2} \sum_{j=1}^{D} \left( \log(\sigma_{k,j}^2) + (x_{ij} - \mu_{k,j})^2 / \sigma_{k,j}^2 \right) \right)$$

where $r_{ik} = E_Q(z_{ik})$. To maximize the sum of the above for all $i$, we separately maximize $\sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \log(\pi_k)$ with respect to $\pi$, and $-\frac{1}{2} \sum_{i=1}^{N} r_{ik} (x_{ij} - \mu_{k,j})^2$ with respect to each $\mu_{k,j}$, and finally $-\frac{1}{2} \sum_{i=1}^{N} r_{ik} \left( \log(\sigma_{k,j}^2) + (x_{ij} - \mu_{k,j})^2 / \sigma_{k,j}^2 \right)$ with respect to each $\sigma_{k,j}^2$. This gives the algorithm presented earlier.