Examples of EM Methods

The method was described and analyzed by Dempster, Laird, and Rubin (1977), although the method had been used much earlier, by Hartley (1958), for example. Many additional details and alternatives are discussed by McLachlan and Krishnan (1997) who also work through about thirty examples of applications of the EM algorithm.

The EM methods can be explained most easily in terms of a random sample that consists of two components, one observed and one unobserved or missing.

A simple example of missing data occurs in life-testing, when, for example, a number of electrical units are switched on and the time when each fails is recorded. In such an experiment, it is usually necessary to curtail the recordings prior to the failure of all units. The failure times of the units still working are unobserved, but the number of censored observations and the time of the censoring obviously provide information about the distribution of the failure times.

Another common example that motivates the EM algorithm is a finite mixture model. This is a more relevant example for data mining. Each observation comes from an unknown one of an assumed set of distributions. The missing data is the distribution indicator. The parameters of the distributions are to be estimated. As a side benefit, the class membership indicator is estimated.

The missing data can be missing observations on the same random variable that yields the observed sample, as in the case of the censoring example; or the missing data can be from a different random variable that is related somehow to the random variable observed.

Many common applications of EM methods do involve missing-data problems, but this is not necessary. Often, an EM method can be constructed based on an artificial "missing" random variable to supplement the observable data.

Example 1

One of the simplest examples of the EM method was given by Dempster, Laird, and Rubin (1977).

Consider the multinomial distribution with four outcomes, that is, the multinomial with probability function,

$$p(x_1, x_2, x_3, x_4) = \frac{n!}{x_1!x_2!x_3!x_4!}\pi_1^{x_1}\pi_2^{x_2}\pi_3^{x_3}\pi_4^{x_4},$$

with $n = x_1 + x_2 + x_3 + x_4$ and $1 = \pi_1 + \pi_2 + \pi_3 + \pi_4$. Suppose the probabilities are related by a single parameter, θ:

$$\begin{aligned}
\pi_1 &= \frac{1}{2} + \frac{1}{4}\theta \\
\pi_2 &= \frac{1}{4} - \frac{1}{4}\theta \\
\pi_3 &= \frac{1}{4} - \frac{1}{4}\theta \\
\pi_4 &= \frac{1}{4}\theta,
\end{aligned}$$

where $0 \leq \theta \leq 1$. (This model goes back to an example discussed by Fisher, 1925, in *Statistical Methods for Research Workers*.)

Given an observation (x_1, x_2, x_3, x_4), the log-likelihood function is

$l(\theta) = x_1 \log(2+\theta) + (x_2+x_3)\log(1-\theta) + x_4 \log(\theta) + \text{constant}.$

Our objective is to estimate θ.

The derivative is

$$\mathrm{d}l(\theta)/\mathrm{d}\theta = \frac{x_1}{2+\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta}.$$

and for this simple problem, the MLE of θ can be determined by solving a simple polynonial equation.

Let's proceed with an EM formulation, however.

To use the EM algorithm on this problem, we can think of a multinomial with five classes, which is formed from the original multinomial by splitting the first class into two with associated probabilities 1/2 and $\theta/4$.

The original variable x_1 is now the sum of u_1 and u_2.

Under this reformulation, we now have a maximum likelihood estimate of θ by considering $u_2 + x_4$ (or $x_2 + x_3$) to be a realization of a binomial with $n = u_2 + x_4 + x_2 + x_3$ and $\pi = \theta$ (or $1 - \theta$).

However, we do not know u_2 (or u_1). Proceeding as if we had a five-outcome multinomial observation with two missing elements, we have the log-likelihood for the complete data,

$$l_c(\theta) = (u_2 + x_4)\log(\theta) + (x_2 + x_3)\log(1 - \theta) + \text{constant},$$

and the maximum likelihood estimate for θ is

$$\frac{u_2 + x_4}{u_2 + x_2 + x_3 + x_4}.$$

The E-step of the iterative EM algorithm fills in the missing or unobservable value with its expected value given a current value of the parameter, $\theta^{(k)}$, and the observed data.

This is a binomial random variable as part of x_1. So with $\theta = \theta^{(k)}$,

$$\begin{aligned} \mathrm{E}_{\theta^{(k)}}(u_2) &= \frac{1}{4}x_1\theta^{(k)} \,/\, \left(\frac{1}{2} + \frac{1}{4}\theta^{(k)}\right) \\ &= u_2^{(k)}. \end{aligned}$$

We now maximize $\mathrm{E}_{\theta^{(k)}}\left(l_c(\theta)\right)$.

Because $l_c(\theta)$ is linear in the data, we have

$$\mathrm{E}\left(l_c(\theta)\right) = \mathrm{E}(u_2 + x_4)\log(\theta) + \mathrm{E}(x_2 + x_3)\log(1-\theta).$$

This maximum occurs at

$$\theta^{(k+1)} = (u_2^{(k)} + x_4)/(u_2^{(k)} + x_2 + x_3 + x_4).$$

The following Matlab statements execute a single iteration.

```
function [u2kp1,tkp1] = em(tk,x)
u2kp1 = x(1)*tk/(2+tk);
tkp1 = (u2kp1 + x(4))/(sum(x)-x(1)+u2kp1);
```

Example 2: A Variation of the Life-Testing Experiment Using an Exponential Model

Consider an experiment described by Flury and Zoppè (2000). It is assumed that the lifetime of light bulbs follows an exponential distribution with mean θ. To estimate θ, n light bulbs were tested until they all failed. Their failure times were recorded as $x_1, \ldots, x_n$. In a separate experiment, m bulbs were tested, but the individual failure times were not recorded. Only the number of bulbs, r, that had failed at time t was recorded.

The missing data are the failure times of the bulbs in the second experiment, $u_1, \ldots, u_m$. We have

$$l_c(\theta\ ;\ x, u) = -n(\log\theta + \bar{x}/\theta) - \sum_{i=1}^{m}(\log\theta + u_i/\theta).$$

The expected value for a bulb still burning is

$$t + \theta$$

and the expected value of one that has burned out is

$$\theta - \frac{t\mathrm{e}^{-t/\theta^{(k)}}}{1 - \mathrm{e}^{-t/\theta^{(k)}}}.$$

Therefore, using a provisional value $\theta^{(k)}$, and the fact that r out of m bulbs have burned out, we have $\mathrm{E}_{U|x,\theta^{(k)}}(l_c)$ as

$$\begin{aligned} q^{(k)}(x,\theta) &= -(n+m)\log\theta \\ &\quad -\frac{1}{\theta}\left(n\bar{x}+(m-r)(t+\theta^{(k)})+r(\theta^{(k)}-th^{(k)})\right), \end{aligned}$$

where $h^{(k)}$ is given by

$$h^{(k)}=\frac{\mathrm{e}^{-t/\theta^{(k)}}}{1-\mathrm{e}^{-t/\theta^{(k)}}}.$$

The k^{th} M step determines the maximum with respect to the variable θ, which, given $\theta^{(k)}$, occurs at

$$\theta^{(k+1)}=\frac{1}{n+m}\left(n\bar{x}+(m-r)(t+\theta^{(k)})+r(\theta^{(k)}-th^{(k)})\right). \tag{1}$$

Starting with a positive number $\theta^{(0)}$, equation (1) is iterated until convergence. The expectation $q^{(k)}$ does not need to be updated explicitly.

To see how this works, let's generate some artificial data and try it out. Some R code to implement this is:

```
# Generate data from an exponential with theta=2, and with the second
# experiment truncated at t=3.  Note that R uses a form of the
# exponential in which the parameter is a multiplier; i.e., the R
# parameter is 1/theta.  Set the seed, so computations are reproducible.
set.seed(4)
n <- 100
m <- 500
theta <- 2
t <- 3
x <- rexp(n,1/theta)
r<-min(which(sort(rexp(m,1/theta))>=3))-1
```

Some R code to implement the EM algorithm:

```
# We begin with  theta=1.
# (Note theta.k is set to theta.kp1 at the beginning of the loop.)
theta.k<-.01
theta.kp1<-1
# Do some preliminary computations.
n.xbar<-sum(x)
# Then loop and test for convergence
   theta.k <- theta.kp1
   theta.kp1 <- (n.xbar +
                 (m-r)*(t+theta.k) +
                 r*(theta.k-
                     t*exp(-t/theta.k)/(1-exp(-t/theta.k))
                   )
                )/(n+m)
```

The value of θ stabilizes to less than 0.1% change at 1.912 in 6 iterations.

This example is interesting because if we assume that the distribution of the light bulbs is uniform, $\mathrm{U}(0, \theta)$ (such bulbs are called “heavybulbs”!), the EM algorithm cannot be applied. Maximum likelihood methods must be used with some care whenever the range of the distribution depends on the parameter. In this case, however, there is another problem. It is in computing $q^{(k)}(x, \theta)$, which does not exist for $\theta < \theta^{(k-1)}$.

Example 3: Estimation in a Normal Mixture Model

A two-component normal mixture model can be defined by two normal distributions, $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$, and the probability that the random variable (the observable) arises from the first distribution is w. The parameter in this model is the vector $\theta = (w, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. (Note that w and the σs have the obvious constraints.)

The pdf of the mixture is

$$p(y; \theta) = wp_1(y; \mu_1, \sigma_1^2) + (1 - w)p_2(y; \mu_2, \sigma_2^2),$$

where $p_j(y; \mu_j, \sigma_j^2)$ is the normal pdf with parameters μ_j and σ_j^2. (I am just writing them this way for convenience; p_1 and p_2 are actually the same parametrized function of course.)

In the standard formulation with $C = (X, U)$, X represents the observed data, and the unobserved U represents class membership. Let $U = 1$ if the observation is from the first distribution and $U = 0$ if the observation is from the second distribution. The unconditional $\mathrm{E}(U)$ is the probability that an observation comes from the first distribution, which of course is w.

Suppose we have n observations on X, $x_1, \ldots, x_n$.

Given a provisional value of θ, we can compute the conditional expected value $\mathrm{E}(U|x)$ for any realization of X. It is merely

$$\mathrm{E}(U|x, \theta^{(k)}) = \frac{w^{(k)} p_1(x; \mu_1^{(k)}, \sigma_1^{2(k)})}{p(x; w^{(k)}, \mu_1^{(k)}, \sigma_1^{2(k)}, \mu_2^{(k)}, \sigma_2^{2(k)})}$$

The M step is just the familiar MLE of the parameters:

$$\begin{aligned}
w^{(k+1)} &= \frac{1}{n} \sum \mathrm{E}(U|x_i, \theta^{(k)}) \\
\mu_1^{(k+1)} &= \frac{1}{nw^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)}) x_i \\
\sigma_1^{2(k+1)} &= \frac{1}{nw^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_1^{(k+1)})^2 \\
\mu_2^{(k+1)} &= \frac{1}{n(1 - w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)}) x_i \\
\sigma_2^{2(k+1)} &= \frac{1}{n(1 - w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_2^{(k+1)})^2
\end{aligned}$$

(Recall that the MLE of σ^2 has a divisor of n, rather than $n - 1$.)

Ingrassia (1992) gives an interesting comparison between the EM approach to this problem and an approach using simulated annealing.

To see how this works, let's generate some artificial data and try it out. Some R code to implement this is:

```
# Normal mixture.   Generate data from nomal mixture with w=0.7,
# mu_1=0, sigma^2_1=1,  mu_2=1, sigma^2_2=2.
# Note that R uses sigma, rather than sigma^2 in rnorm.
#  Set the seed, so computations are reproducible.
set.seed(4)
n <- 300
w <- 0.7
mu1 <- 0
sigma21 <- 1
mu2 <- 5
sigma22 <- 2
x <- ifelse(runif(n)<w,
rnorm(n,mu1,sqrt(sigma21)),rnorm(n,mu2,sqrt(sigma22)))
```

First, assume that μ_1, σ_1^2, μ_2, and σ_2^2 are all known:

```
# Initialize.
theta.k<-.1
theta.kp1<-.5

# Then loop over the following
   theta.k <- theta.kp1
   tmp <- theta.k*dnorm(x, mu1,sqrt(sigma21))
   ehat.k <- tmp/(tmp+(1-theta.k)*dnorm(x, mu2,sqrt(sigma22)))
   theta.kp1<- mean(ehat.k)
```

This converges very quickly to 0.682, at which point the parameter estimate changes less than 0.1%.

I next tried the case where only σ_1^2 and σ_2^2 are assumed known. This did not converge, but I didn't figure out why not.