

The EM Method

Introduction

We will assume a problem of estimating the parameter θ (which may, of course, be a vector). We will assume a likelihood function $L(\theta; y)$ of known form. In MLE, the likelihood is the objective function, and we seek to maximize it.

Statistical estimation based on optimizing an objective function usually involves an iterative method. (A quadratic objective function is a notable exception to this statement.) In an iterative method, we move through a sequence $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots$ that converges to the optimal value.

The computational burden in a single iteration for solving the MLE optimization problem can be reduced by more than a linear amount by separating θ into two subvectors. The MLE is then computed by alternating between computations involving the two subvectors, and the iterations proceed in a zigzag path to the solution. Each of the individual sequences of iterations is simpler than the sequence of iterations on the full θ .

Another alternating method that arises from an entirely different approach alternates between updating $\theta^{(k)}$ using maximum likelihood and conditional expected values. This method is called the *EM method* because the alternating steps involve an expectation and a maximization.

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.

Description

Let $Y = (U, V)$, and assume that we have observations on U but not on V . There are thus two likelihoods, one based on the complete (but unknown) sample, and one based only on the observed sample. We wish to estimate the parameter θ , which figures in the distribution of both components of Y .

Because we do not have all the observations we cannot write the likelihood exactly. We can use a provisional value of $\theta^{(k)}$ to approximate the complete likelihood based on the expected value of V . The approximation will generally now be a function of both θ and $\theta^{(k)}$. We then maximize this approximation with respect to θ to get $\theta^{(k+1)}$.

Let $L_c(\theta ; u, v)$ and $l_{L_c}(\theta ; u, v)$ denote, respectively, the likelihood and the log-likelihood for the complete sample. The likelihood for the observed U is

$$L(\theta ; u) = \int L_c(\theta ; u, v) dv,$$

and $l_L(\theta ; u) = \log L(\theta ; u)$. This is a function we can maximize, if the maximum exists. The problem, however, is to determine this function; that is, to average over V . (This is what the integral is doing, but we do not know what to integrate.) The average over V is the expected value with respect to the marginal distribution of V . This is a standard problem, and we estimate the expectation using the observations on U and a provisional value of θ .

We begin with a provisional value of θ , call it $\theta^{(0)}$. Given any provisional value $\theta^{(k)}$, we will compute a provisional value $\theta^{(k+1)}$ that does not decrease the likelihood.

The EM approach to maximizing $L(\theta ; u)$ has two alternating steps. The steps are iterated until convergence.

- E step : compute $q^{(k)}(u, \theta^{(k)}) = E_{V|u, \theta^{(k)}}(l_{L_c}(\theta | u, V))$.
- M step : determine $\theta^{(k+1)}$ to maximize $q^{(k)}(u, \theta^{(k)})$, subject to any constraints on acceptable values of θ .

The sequence $\theta^{(1)}, \theta^{(2)}, \dots$ converges to a local maximum of the observed-data likelihood $L(\theta ; u)$ under fairly general conditions (including, of course, the nonexistence of a local maximum near enough to $\theta^{(0)}$).

The EM method can be very slow to converge, however. See Wu (1983) for discussion of the convergence conditions.

Example

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

$$l(\theta) = x_1 \log(2 + \theta) + (x_2 + x_3) \log(1 - \theta) + x_4 \log(\theta) + c$$

and

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

The objective is to estimate θ .

Dempster, Laird, and Rubin used $n = 197$ and $x = (125, 18, 20, 34)$. For this simple problem, the MLE of θ can be determined by solving a simple polynomial equation.

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 x_{11} and x_{12} . Under this reformulation, we now have a maximum likelihood estimate of θ by considering $x_{12} + x_4$ (or $x_2 + x_3$) to be a realization of a binomial with $n = x_{12} + x_4 + x_2 + x_3$ and $\pi = \theta$ (or $1 - \theta$). However, we do not know x_{12} (or x_{11}). Proceeding as if we had a five-outcome multinomial observation with two missing elements, we have the log-likelihood for the complete data,

$$l_c(\theta) = (x_{12} + x_4) \log(\theta) + (x_2 + x_3) \log(1 - \theta),$$

and the maximum likelihood estimate for θ is

$$\frac{x_{12} + x_4}{x_{12} + x_2 + x_3 + x_4}.$$

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

$$E(l_c(\theta)) = E(x_{12} + x_4) \log(\theta) + E(x_2 + x_3) \log(1 - \theta).$$

Under this setup, with $\theta = \theta^{(k)}$,

$$\begin{aligned} E_{\theta^{(k)}}(x_{12}) &= \frac{1}{4}x_1\theta^{(k)} / \left(\frac{1}{2} + \frac{1}{4}x_1\theta^{(k)}\right) \\ &= x_{12}^{(k)}. \end{aligned}$$

We now maximize $E_{\theta^{(k)}}(l_c(\theta))$. This maximum occurs at

$$\theta^{(k+1)} = (x_{12}^{(k)} + x_4) / (x_{12}^{(k)} + x_2 + x_3 + x_4).$$

The following Matlab statements execute a single iteration.

```
function [x12kp1,tkp1] = em(tk,x)
x12kp1 = x(1)*tk/(2+tk);
tkp1 = (x12kp1 + x(4))/(sum(x)-x(1)+x12kp1);
```

Another Example: 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 u_1, \dots, u_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, v_1, \dots, v_m . We have

$$l_{L_c}(\theta ; u, v) = -n(\log \theta + \bar{u}/\theta) - \sum_{i=1}^m (\log \theta + v_i/\theta).$$

The expected value for a bulb still burning is

$$t + \theta$$

and the expected value of one that has burned out is

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

Therefore, using a provisional value $\theta^{(k)}$, and the fact that r out of m bulbs have burned out, we have $E_{V|u,\theta^{(k)}}(l_{L_c})$ as

$$q^{(k)}(u, \theta) = -(n + m) \log \theta - \frac{1}{\theta} \left(n\bar{u} + (m - r)(t + \theta^{(k)}) + r(\theta^{(k)} - th^{(k)}) \right),$$

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

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

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

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

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 S-Plus code to implement this is:

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

Some S-Plus 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.ubar<-sum(u)
# Then loop and test for convergence
theta.k <- theta.kp1
theta.kp1 <- (n.ubar +
              (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, $U(0, \theta)$ (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 $q^{(k)}(u, \theta)$, which does not exist for $\theta < \theta^{(k-1)}$.

Another Example: Estimation in a Normal Mixture Model

A two-component normal mixture model can be defined by two normal distributions, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, and the statement 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 σ s 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 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 $Y = (U, V)$, U represents the observed data, and the unobserved V represents class membership. Let $V = 1$ if the observation is from the first distribution and $V = 0$ if the observation is from the second distribution. The unconditional $E(V)$ is the probability that an observation comes from the first distribution, which of course is w .

Suppose we have n observations on U , u_1, \dots, u_n .

Given a provisional value of θ , we can compute the conditional expected value $E(V|u)$ for any realization of U . It is merely

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

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

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

(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 S-Plus code to implement this is:

```
# Normal mixture.  Generate data from normal mixture with w=0.7,
# mu_1=0, sigma^2_1=1, mu_2=1, sigma^2_2=2.
# Note that S-Plus uses a 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
u <- 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(u, mu1,sqrt(sigma21))
  ehat.k <- tmp/(tmp+(1-theta.k)*dnorm(u, 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.

Alternative Ways of Performing the Computations

There are two kinds of computations that must be performed in each iteration:

- E step : compute $q^{(k)}(u, \theta^{(k-1)}) = E_{V|u, \theta^{(k-1)}}(l_{L_c}(\theta | u, V))$.
- M step : determine $\theta^{(k)}$ to maximize $q^{(k)}(u, \theta^{(k-1)})$, subject to any constraints on acceptable values of θ .

Although in the paper that first provided a solid description of the EM method (Dempster, Laird, and Rubin, 1977), specific techniques were used for the computations in the two steps, it is not necessary for the EM method to use those same inner-loop algorithms. There are various other ways to perform each of these computations.

A number of papers since 1977 have suggested specific methods for the computations and have given new names to methods based on those inner-loop computations.

E Step

There are different ways the expectation step can be carried out. In the happy case of an exponential family or some other nice distributions, the expectation can be computed in closed form. Otherwise, computing the expectation is a numerical quadrature problem. There are various procedures for quadrature, including Monte Carlo. Wei and Tanner (1990) call an EM method that uses Monte Carlo to evaluate the expectation an MCEM method. (If a Newton-Cotes method is used, however, we do not call it an NCEM method.) The additional Monte Carlo computations add a lot to the overall time required for convergence of the EM method. In order to reduce the costs of Monte Carlo sampling in MCEM, Levine and Casella (2001) suggest reuse of Monte Carlo samples from previous expectation steps.

An additional problem in using Monte Carlo in the expectation step may be that the distribution of Y is difficult to simulate. The convergence criterion for optimization methods that involve Monte Carlo generally should be tighter than for deterministic methods.

M Step

For the maximization step, there are more choices, as we have seen in the discussion of maximum likelihood estimation above.

For the maximization step, Dempster, Laird, and Rubin (1977) suggested requiring only an increase in the expected value; that is, take $\theta^{(k)}$ so that $q_k(u, \theta^{(k)}) \geq q_{k-1}(u, \theta^{(k-1)})$. This is called a generalized EM algorithm, or GEM. Rai and Matthews (1993) suggest taking $\theta^{(k)}$ as the point resulting from a single Newton step and called this method EM1.

Meng and Rubin (1993) describe a GEM algorithm in which the M-step is a componentwise maximization; that is, if $\theta = (\theta_1, \theta_2)$, first $\theta_1^{(k)}$ is determined to maximize q subject to the constraint $\theta_2 = \theta_2^{(k-1)}$; then $\theta_2^{(k)}$ is determined to maximize q subject to the constraint $\theta_1 = \theta_1^{(k)}$. They call this an expectation conditional maximization, or ECM, algorithm. This sometimes simplifies the maximization problem so that it can be done in closed form. Jamshidian and Jennrich (1993) discuss acceleration of the EM algorithm using conjugate gradient methods and using quasi-Newton methods (Jamshidian and Jennrich, 1997).

Kim and Taylor (1995) describe an EM method when there are linear restrictions on the parameters.

Alternate Ways of Terminating the Computations

In any iterative algorithm, we must have some way of deciding to terminate the computations. (The generally-accepted definition of “algorithm” requires that it terminate. In any event, of course, we want the computations to cease at some point.)

One way of deciding to terminate the computations is based on convergence; if the computations have converged we quit. In addition, we also have some criterion by which we decide to quit anyway.

In an iterative optimization algorithm, there are two obvious ways of deciding when convergence has occurred. One is when the decision variables (the estimates in MLE) are no longer changing appreciably, and the other is when the value of the objective function (the likelihood) is no longer changing appreciably.

It is easy to think of cases in which the objective function converges, but the decision variables do not. All that is required is that the objective function is flat over a region at its maximum. In statistical terms, this corresponds to unidentifiability.

See Wu (1983) for discussion of the convergence conditions.

The Variance of Estimators Defined by the EM Method

As is usual for estimators defined as solutions to optimization problems, we may have some difficulty in determining the statistical properties of the estimators.

Louis (1982) suggested a method of estimating the variance-covariance matrix of the estimator by use of the gradient and Hessian of the complete-data log-likelihood, $l_{L_c}(\theta ; u, v)$.

Meng and Rubin (1991) use a “supplemented” EM method, SEM, for estimation of the variance-covariance matrix.

Kim and Taylor (1995) also described ways of estimating the variance-covariance matrix using computations that are part of the EM steps.

It is interesting to note that under certain assumptions on the distribution, the iteratively reweighted least squares method can be formulated as an EM method (see Dempster, Laird, and Rubin, 1980).