The EM Algorithm

Kenneth Lange

Departments of Biomathematics
and Human Genetics

David Geffen School of Medicine at UCLA

May, 2003

Overview of the EM Algorithm

1. Maximum likelihood estimation is ubiquitous in statistics

2. EM is a special case of the MM algorithm that relies on the notion of missing information.

3. The surrogate function is created by calculating a certain conditional expectation. Sometimes an MM and an EM algorithm coincide for the same problem; sometimes not.

4. Convexity enters through Jensen's inequality.

5. Many examples were known before the general principle was enunciated.

Nature of Missing Information

1. Missing information can take the form of missing data.

2. Missing information can also be more abstract. Even with perfect data collection, there can be missing information.

3. For instance, in PET scanning, current machines cannot determine where along a projection line a decay event has taken place. If we knew the pixel of origin of each decay event, then estimating the concentration of a radio-labeled compound would be straightforward.

4. The complete data should be conceptualized to make maximum likelihood estimation trivial.

Ingredients of the EM Algorithm

1. The observed data y with likelihood $f(y \mid \theta)$. Here θ is a parameter vector.

2. The complete data x with likelihood $g(x \mid \theta)$.

3. The conditional expectation

$$Q(\theta \mid \theta^n) \;=\; \mathrm{E}[\ln g(x \mid \theta) \mid y, \theta^n]$$

   furnishes the minorizing function up to a constant. Here θ^n is the value of θ at iteration n of the EM algorithm.

4. Calculation of $Q(\theta \mid \theta^n)$ constitutes the E step; maximization of $Q(\theta \mid \theta^n)$ with respect to θ constitutes the M step.

Minorization Property of the EM Algorithm

1. The proof depends on Jensen's inequality $\mathsf{E}[h(Z)] \geq h[\mathsf{E}(Z)]$ for a random variable Z and convex function $h(z)$.

2. If $p(z)$ and $q(z)$ are probability densities with respect to a measure μ, then the convexity of $-\ln z$ implies the information inequality

   $$\mathsf{E}_p[\ln p] - \mathsf{E}_p[\ln q] = \mathsf{E}_p[-\ln \frac{q}{p}] \geq -\ln \mathsf{E}_p(\frac{q}{p}) = -\ln \int \frac{q}{p} p \, d\mu = 0,$$

   with equality when $p = q$.

3. In the E step minorization, we apply the information inequality to the conditional densities $p(x) = f(x \mid \theta^n)/g(y \mid \theta^n)$ and $q(x) = f(x \mid \theta)/g(y \mid \theta)$ of the complete data x given the observed data y.

Minorization Property II

1. The information inequality $\mathsf{E}_p[\ln p] \geq \mathsf{E}_p[\ln q]$ now yields

$$Q(\theta \mid \theta^n) - \ln g(y \mid \theta) = \mathsf{E}\left[\ln \frac{f(x \mid \theta)}{g(y \mid \theta)} \mid y, \theta^n\right]$$

$$\leq \mathsf{E}\left[\ln \frac{f(x \mid \theta^n)}{g(y \mid \theta^n)} \mid y, \theta^n\right] = Q(\theta^n \mid \theta^n) - \ln g(y \mid \theta^n),$$

with equality when $\theta = \theta^n$.

2. Thus, $Q(\theta \mid \theta^n) - Q(\theta^n \mid \theta^n) + \ln g(y \mid \theta^n)$ minorizes $\ln g(y \mid \theta)$.

3. In the M step it suffices to maximize $Q(\theta \mid \theta^n)$ since the other two terms of the minorizing function do not depend on θ.

Example 1: Twin Data

1. You are given a sample of m male twin pairs, f female twin pairs, and o opposite sex twin pairs. Estimate the probability p that a twin pair is identical and the probability q that a child is male.

2. Here $y = (m, f, o)$ is the observed data and $\theta = (p, q)$ is the parameter vector. If we knew exactly which pairs of same-sex twins were identical, then it would be easy to estimate p and q. Thus, we postulate complete data $x = (m_1, m_2, f_1, f_2, o)$, with m_1 representing the number of male identical twin pairs and m_2 the number of male non-identical twin pairs. f_1 and f_2 are defined similarly.

Example 1: Complete Data Loglikelihood

1. The multinomial complete data likelihood is

$$\begin{aligned} g(x \mid \theta) &= \binom{m+f+o}{m_1,\ m_2,\ f_1,\ f_2,\ o} (pq)^{m_1}[(1-p)q^2]^{m_2}[p(1-q)]^{f_1} \\ &\quad \times [(1-p)(1-q)^2]^{f_2}[(1-p)2q(1-q)]^{o} \end{aligned}$$

since identical twins involve one choice of sex and non-identical twins two choices of sex.

2. The complete data loglikelihood is

$$\begin{aligned} \ln g(x \mid \theta) &= (m_1+f_1)\ln p + (m_2+f_2+o)\ln(1-p) \\ &\quad +(m_1+2m_2+o)\ln q \\ &\quad +(f_1+2f_2+o)\ln(1-q) + \text{constant.} \end{aligned}$$

Example 1: E Step

To carry out the E step we calculate

$$
\begin{aligned}
m_1^n = \mathsf{E}(m_1 \mid y, \theta^n) &= m \frac{p^n q^n}{p^n q^n + (1-p^n)(q^n)^2} \\
m_2^n = \mathsf{E}(m_2 \mid y, \theta^n) &= m \frac{(1-p^n)(q^n)^2}{p^n q^n + (1-p^n)(q^n)^2} \\
f_1^n = \mathsf{E}(f_1 \mid y, \theta^n) &= f \frac{p^n(1-q^n)}{p^n(1-q^n) + (1-p^n)(1-q^n)^2} \\
f_2^n = \mathsf{E}(f_2 \mid y, \theta^n) &= f \frac{(1-p^n)(1-q^n)^2}{p^n(1-q^n) + (1-p^n)(1-q^n)^2}
\end{aligned}
$$

by applying Bayes' rule.

Example 1: M Step

1. The surrogate function is

$$\begin{aligned} Q(\theta \mid \theta^n) &= (m_1^n + f_1^n)\ln p + (m_2^n + f_2^n + o)\ln(1-p) \\ &\quad + (m_1^n + 2m_2^n + o)\ln q \\ &\quad + (f_1^n + 2f_2^n + o)\ln(1-q) + \text{constant}. \end{aligned}$$

2. Straightforward calculus shows the maximum occurs for

$$\begin{aligned} p^{n+1} &= \frac{m_1^n + f_1^n}{m + f + o} \\ q^{n+1} &= \frac{m_1^m + 2m_2^n + o}{m + f + o + m_2^n + f_2^n}. \end{aligned}$$

   Note that $Q(\theta \mid \theta^n)$ is separable in the parameters p and q and that $m_1 + m_2 + f_1 + f_2 + o = m + f + o$.

Example 1: Hidden Binomial Updates

1. Both the number of identical twins and the number of choices of the male sex involve hidden binomial trials.

2. The update in such circumstances take the form

$$r^{n+1} = \frac{\mathsf{E}(\#\text{successes} \mid y, \theta^n)}{\mathsf{E}(\#\text{trials} \mid y, \theta^n)}$$

   for $r = p$ or $r = q$. In the first case the number of trials is fixed at $m + f + o$, and in the second case the number of trials is random because the number of choices of sex depends on the number of identical twins versus the number of non-identical twins.

Example 2: Light Bulb Lifetimes

1. The random lifetime of a light bulb is postulated to be exponential with unknown mean $1/\theta$.

2. The lifetimes $y_1, \ldots, y_r$ of r independent bulbs are observed. A further s independent bulbs are observed at time $t > 0$. Bulb $i + r$ is registered as still burning, $z_{i+r} = 1$, or expired, $z_{i+r} = 0$. Thus, the lifetimes of the second set of bulbs are both left and right censored.

3. In this situation it is natural to view the complete data as the observed lifetimes $y_1, \ldots, y_r$ supplement by the unobserved lifetimes $x_{r+1}, \ldots, x_{r+s}$.

Example 2: Complete Data Loglikelihood

The complete data loglikelihood is

$$\begin{aligned} g(x \mid \theta^n) &= \sum_{i=1}^{r} (\ln\theta - \theta y_i) + \sum_{i=r+1}^{r+s} (\ln\theta - \theta x_i) \\ &= r\ln\theta - r\theta\bar{y} + \sum_{i=r+1}^{r+s} (\ln\theta - \theta x_i) \end{aligned}$$

for exponentially distributed lifetimes, where $\bar{y}$ is the average value of $y_1, \ldots, y_r$.

Example 2: E Step

1. Because the survival time of a light bulb lacks memory, right censored data gives $\mathrm{E}(x_{r+i} \mid z_{r+i} = 1, \theta^n) = t + 1/\theta^n$.

2. For left censored data, integration by parts and the fundamental theorem of calculus yield

$$\begin{aligned} \mathrm{E}(x_{r+i} \mid z_{r+i} = 0, \theta^n) &= \frac{\int_0^t x\theta^n e^{-\theta^n x} dx}{\int_0^t \theta^n e^{-\theta^n x} dx} \\ &= \frac{1}{\theta^n} - \frac{te^{-\theta^n t}}{1 - e^{-\theta^n t}}. \end{aligned}$$

3. Weighted by their respective probabilities and summed, these two conditional expectations give back the unconditional mean $1/\theta^n$ of x_{r+i}.

Example 2: M Step

1. If we let $\mu_i^n = \mathrm{E}(x_{r+i} \mid z_{r+i}, \theta^n)$, then the surrogate function is

$$Q(\theta \mid \theta^n) \;\; = \;\; (r+s)\ln\theta - \theta[r\bar{y} + \sum_{i=r+1}^{r+s} \mu_i^n].$$

2. It is now easy to differentiate and solve for the EM update

$$\theta^{n+1} \;\; = \;\; \frac{r+s}{r\bar{y} + \sum_{i=r+1}^{r+s} \mu_i^n}.$$

3. In other words, we fill in unknown times by their conditional expectations and then identity $1/\theta^{n+1}$ with the average of the actual and imputed lifetimes.

Example 3: Binomial-Poisson Mixture

Thisted considers historic data on widows and their dependent children from a Swedish pension fund. If y_k denotes the number of widows with k children, then the data values are $y_0 = 3062$, $y_1 = 587$, $y_2 = 284$, $y_3 = 103$, $y_4 = 33$, and $y_5 = 4$, and $y_6 = 2$. The fact that most widows have no dependent children suggests that a simple Poisson model would give a poor fit. A better model is a mixture of a population of widows with no children, population A, and a population of widows having a Poisson number of children, population B. Suppose a widow falls into population A with probability p and into population B with probability $1 - p$. Let μ be the mean of the Poisson distribution characterizing population B.

Example 3: Loglikelihood for the Binomial-Poisson Mixture

1. The parameter vector is $\theta = (p, \mu)$.

2. Omitting the obvious multinomial coefficient, the loglikelihood for the observed data is

$$\begin{aligned} \ln g(y \mid \theta) &= y_0 \ln[p + (1-p)e^{-\mu}] \\ &\quad + \sum_{k \geq 1} y_k [\ln(1-p) + k \ln \mu - \mu - \ln k!] \end{aligned}$$

3. There is no closed-form maximum.

Example 3: The Complete Data

1. To generate the complete data, we split the y_0 widows into x_A widows from population A and x_B widows from population B.

2. The loglikelihood for the complete data is

$$\begin{aligned} \ln f(x \mid \theta) &= x_A \ln p + x_B[\ln(1-p) - \mu] \\ &\quad + \sum_{k \geq 1} y_k[\ln(1-p) + k \ln \mu - \mu - \ln k!] \end{aligned}$$

3. In the E step, we calculate

$$\begin{aligned} x_A^n = \mathrm{E}(x_A \mid y_0, \theta^n) &= y_0 \frac{p^n}{p^n + (1-p^n)e^{-\mu^n}} \\ x_B^n = \mathrm{E}(x_B \mid y_0, \theta^n) &= y_0 - \mathrm{E}(x_A \mid y_0, \theta^n). \end{aligned}$$

Example 3: The M Step

1. The surrogate function is

$$\begin{aligned} Q(\theta \mid \theta^n) &= x_A^n \ln p + x_B^n[\ln(1-p) - \mu] \\ &\quad + \sum_{k \geq 1} y_k[\ln(1-p) + k \ln \mu - \mu - \ln k!] \end{aligned}$$

2. The maximum occurs for

$$\begin{aligned} p^{n+1} &= \frac{x_A^n}{y_0 + \sum_{k \geq 1} y_k} \\ \mu^{n+1} &= \frac{\sum_{k \geq 1} k y_k}{x_B^n + \sum_{k \geq 1} y_k}. \end{aligned}$$

Example 3: The Algorithm in Practice

1. The updates are hidden binomial and hidden Poisson updates.

2. The first few iterations are:

$$\begin{array}{ll} p^0 = 0.75000 & \mu^0 = 0.40000 \\ p^1 = 0.61418 & \mu^1 = 1.03548 \\ p^2 = 0.61438 & \mu^2 = 1.03601 \\ p^3 = 0.61453 & \mu^3 = 1.03643 \\ p^4 = 0.61465 & \mu^4 = 1.03675 \\ p^4 = 0.61474 & \mu^5 = 1.03670 \end{array}$$

3. Convergence is fast at first and then slows.

Example 4: Transmission Tomography

1. Recall that the observed data consist of the photon counts y_i for the various projection lines i. Along projection i the number of photons that begin the journey from source to detector follows a Poisson distribution with mean d_i. Pixel j is assigned attenuation coefficient θ_j, and projection i intersects pixel j over a distance of l_{ij}.

2. With this notation the loglikelihood of the observed data is

$$\ln g(y \mid \theta) = \sum_i [-d_i e^{-\sum_j l_{ij}\theta_j} - y_i \sum_j l_{ij}\theta_j + y_i \ln d_i - \ln y_i!].$$

Example 4: Complete Data

1. The complete data consist of the number of photons that enter pixel j along projection i for all pairs i and j.

2. Since transmission acts independently along each projection, we focus on a single projection and drop the projection subscript. Let $y = x_m$ be the number of photons detected and x_j the number of photons entering pixel j. Here we assume $m - 1$ pixels along the projection.

3. Each of these random variables is Poisson; x_1 is Poisson by virtue of how X-rays are generated, and x_j is Poisson because random thinning turns one Poisson process into another.

Example 4: Complete Data Likelihood

1. For the sake of simplicity, we omit a smoothing prior.

2. Given x_j, the number of photons x_{j+1} passing through pixel j is binomial with mean x_j and success probability $e^{-l_j\theta_j}$.

3. The complete data loglikelihood is therefore

$$\begin{aligned} f(x \mid \theta) &= -d + x_1 \ln d - \ln x_1! + \sum_{j=1}^{m-1} \Big[\ln \binom{x_j}{x_{j+1}} \\ &\quad + x_{j+1} \ln e^{-l_j\theta_j} + (x_{j+1} - x_j) \ln(1 - e^{-l_j\theta_j}) \Big]. \end{aligned}$$

Example 4: E Step

1. To complete the E step, it suffices to calculate $\mathrm{E}(x_j \mid x_m, \theta^n)$ for each j.

2. The unconditional mean $\mu_j = \mathrm{E}(x_j) = de^{-\sum_{k=1}^{j-1} l_k \theta_k}$.

3. On the next slide we show that $\mathrm{E}(x_j \mid x_m, \theta^n) = \mu_j - \mu_m + x_m$.

4. This will show in the original notation that

$$Q(\theta \mid \theta^n) \;=\; \sum_i \sum_j [-r_{ij}^n l_{ij} \theta_j + (s_{ij}^n - r_{ij}^n) \ln(1 - e^{-l_{ij}\theta_j})]$$

   for computable constants r_{ij}^n and s_{ij}^n depending on θ^n and the y_i.

Example 4: Calculation of $\mathsf{E}(x_j \mid x_m, \theta^n)$

Suppose U and V are Poisson counts. If V is generated from U by randomly thinning each U point with probability $1-p$, then $U-V$ is Poisson and independent of V. If U has mean μ, then

$$\begin{aligned} \Pr(U-V=j \mid V=k) &= \frac{\frac{\mu^{j+k}}{(j+k)!}e^{-\mu}\binom{j+k}{k}p^k(1-p)^j}{\frac{(p\mu)^k}{(k)!}e^{-p\mu}} \\ &= \frac{[(1-p)\mu]^j}{j!}e^{-(1-p)\mu}. \end{aligned}$$

Thus, $\mathsf{E}(U-V \mid V) = (1-p)\mu$ and $\mathsf{E}(U \mid V) = V + (1-p)\mu$. Now apply this to $V = x_m$ and $U = x_j$.

Example 4: M Step

1. The surrogate function

$$Q(\theta \mid \theta^n) \;=\; \sum_i \sum_j [-r_{ij}^n l_{ij}\theta_j + (s_{ij}^n - r_{ij}^n)\ln(1 - e^{-l_{ij}\theta_j})]$$

   separates the parameters.

2. To find the maximum of the part containing θ_j, one must solve a transcendental equation numerically.

3. This is not hard, but the EM algorithm is inferior to the MM algorithm posed earlier because of the work involved in computing the constants r_{ij}^n and s_{ij}^n. In essence, one must exponentiate all of the partial line integrals running from the source to each intermediate pixel along each projection.

Concluding Comments on the EM Algorithm

1. It always involves missing information. Recognizing an appropriate complete data framework is often fairly natural.

2. The E step can involve tricky conditional expectations. Never guess at the form of the surrogate. Work through the recipe.

3. Convergence can be very slow on some problems and is intimately related to the amount of missing information.

4. Intermediate quantities in the algorithm often have useful statistical interpretations.

5. Every EM algorithm is an MM algorithm, so all convergence results carry over.

References

1. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J Roy Stat Soc B* 39:1–38

2. Lange K (1999) *Numerical Analysis for Statisticians.* Springer-Verlag, New York

3. Little RJA, Rubin DB (1987) *Statistical Analysis with Missing Data*. Wiley, New York

4. McLachlan GJ, Krishnan T (1996) *The EM Algorithm and Extensions*.