Optimization Methods II.
EM algorithms. Exercises.

Anatoli Iambartsev
IME-USP

[RC] Example 5.16.

A classic example of the EM algorithm is a genetics problem [DLR] where observations (x_1, x_2, x_3, x_4) are gathered from the multinomial distribution

$$(x_1, x_2, x_3, x_4) \sim M\Big(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\Big)$$

with $n = x_1 + x_2 + x_3 + x_4$. Thus the observed likelihood

$$L(\theta \mid x_1, x_2, x_3, x_4) \propto (2+\theta)^{x_1}\theta^{x_4}(1-\theta)^{x_2+x_3}.$$

[RC] Example 5.16.

Estimation is easier if the x_1 cell is split into two cells, so we create the augmented model

$$(z_1, z_2, x_2, x_3, x_4) \sim M\Big(n; \frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\Big),$$

with $x_1 = z_1 + z_2$. Thus the complete likelihood

$$L^c(\theta \mid z_1, z_2, x_2, x_3, x_4) \propto \theta^{z_2+x_4}(1-\theta)^{x_2+x_3}.$$

Note that

$$Z_2 \mid x_1 \sim B\Big(x_1, \frac{\theta}{\theta+2}\Big) \text{ and } \mathbb{E}_\theta(Z_2 \mid x_1) = \frac{\theta}{\theta+2}x_1.$$

[RC] Example 5.16.

The expected complete log-likelihood function is

$$\begin{aligned}&\mathbb{E}_{\theta_0}\big((Z_2 + x_4)\log\theta + (x_2 + x_3)\log(1-\theta)\big)\\ &\qquad = \Big(\frac{\theta_0}{\theta_0 + 2}x_1 + x_4\Big)\log\theta + (x_2 + x_3)\log(1-\theta),\end{aligned}$$

which can easily be maximized in θ, leading to the EM step

$$\hat{\theta}_1 = \Big\{\frac{\theta_0 x_1}{2+\theta_0} + x_4\Big\} \Big/ \Big\{\frac{\theta_0 x_1}{2+\theta_0} + x_2 + x_3 + x_4\Big\}.$$

[RC] Example 5.16.

A Monte Carlo EM solution would replace the expectation $\theta_0 x_1/(2+\theta_0)$ with the empirical average

$$\bar{z}_m = \frac{1}{m}\sum_{i=1}^{m} z_i,$$

where the z_i are simulated from a binomial distribution $B(x_1, \theta_0/(2+\theta_0))$, or, equivalently, by

$$m\bar{z}_m \sim B(mx_1, \theta_0/(2+\theta_0)).$$

The MCEM step would then be

$$\hat{\hat{\theta}} = \frac{\bar{z}_m + x_4}{\bar{z}_m + x_2 + x_3 + x_4} \to \hat{\theta},$$

when m grows to infinity.

This example is merely a formal illustration of the Monte Carlo EM algorithm and its convergence properties since EM can be applied.

[RC] Example 5.17. The next example, however, details a situation in which the E-step is too complicated to be implemented and where the Monte Carlo EM algorithm provides a realistic (if not straightforward) alternative.

A simple random effect logit model processed in Booth and Hobert (1999) represents observations $y_{ij} (i = 1, ..., n, j = 1, ..., m)$ as distributed conditionally on one covariate x_{ij} as a logit model

$$P(y_{ij} = 1 \mid x_{ij}, u_i, \beta) = \frac{\exp(\beta x_{ij} + u_i)}{1 + \exp(\beta x_{ij} + u_i)},$$

where $u_i \sim N(0, \sigma^2)$ is an unobserved random effect. The vector of random effects $(U_1, ..., U_n)$ therefore corresponds to the missing data $\mathbf{Z}$. When considering $Q(\theta' \mid \theta, \mathbf{x}, \mathbf{y})$, with $\theta = (\beta, \sigma)$

$$\begin{aligned} Q(\theta' \mid \theta, \mathbf{x}, \mathbf{y}) \quad = \quad & \sum_{i,j} y_{ij} \mathbb{E}\Big(\beta' x_{ij} + U_i \mid \beta, \sigma, \mathbf{x}, \mathbf{y}\Big) \\ & - \sum_{i,j} \mathbb{E}\Big(\log(1 + \exp(\beta' x_{ij} + U_i)) \mid \beta, \sigma, \mathbf{x}, \mathbf{y}\Big) \\ & - \sum_{i} \mathbb{E}\Big(U_i^2 \mid \beta, \sigma, \mathbf{x}, \mathbf{y}\Big) / 2(\sigma')^2 - n \log \sigma', \end{aligned}$$

[RC] Example 5.17. When considering $Q(\theta' \mid \theta, \mathbf{x}, \mathbf{y})$, with $\theta = (\beta, \sigma)$

$$\begin{aligned} Q(\theta' \mid \theta, \mathbf{x}, \mathbf{y}) &= \sum_{i,j} y_{ij} \mathbb{E}\Big(\beta' x_{ij} + U_i \mid \beta, \sigma, \mathbf{x}, \mathbf{y}\Big) \\ &\quad - \sum_{i,j} \mathbb{E}\Big(\log(1 + \exp(\beta' x_{ij} + U_i)) \mid \beta, \sigma, \mathbf{x}, \mathbf{y}\Big) \\ &\quad - \sum_{i} \mathbb{E}\Big(U_i^2 \mid \beta, \sigma, \mathbf{x}, \mathbf{y}\Big)/2(\sigma')^2 - n \log \sigma', \end{aligned}$$

it is impossible to compute the expectations in U_i. Were those available, the M-step would then be almost straightforward since maximizing $Q(\theta' \mid \theta, \mathbf{x}, \mathbf{y})$ in σ' leads to

$$(\sigma')^2 = \frac{1}{n} \sum_i \mathbb{E}\Big(U_i^2 \mid \beta, \sigma, \mathbf{x}, \mathbf{y}\Big)$$

maximizing $Q(\theta' \mid \theta, \mathbf{x}, \mathbf{y})$ in β' produces the fixed-point equation

$$\sum_{i,j} y_{ij} x_{ij} = \sum_{i,j} x_{ij} \mathbb{E}\left(\frac{\exp(\beta' x_{ij} + U_i)}{1 + \exp(\beta' x_{ij} + U_i)} \middle| \beta, \sigma, \mathbf{x}, \mathbf{y} \right)$$

which is not particularly easy to solve in β.

[RC] Example 5.17.

The alternative to EM is therefore to simulate the U_i's conditional on $\beta, \sigma, \mathbf{x}, \mathbf{y}$ in order to replace the expectations above with Monte Carlo approximations. While a direct simulation from

$$\pi(u_i \mid \beta, \sigma, \mathbf{x}, \mathbf{y}) \propto \frac{\exp\left\{\sum_j y_{ij} u_i - u_i^2/2\sigma^2\right\}}{\prod_j \left(1 + \exp(\beta x_{ij} + u_i)\right)}$$

is feasible ([BH] Booth and Hobert, 1999), it requires some preliminary tuning better avoided at this stage, and it is thus easier to implement an MCMC version of the simulation of the u_i's toward the approximations of both expectations.

[FZ] First Exercise.

“Suppose that the lifetime of litebulbs follows an exponential distribution with unknown mean θ. A total of $M + N$ litebulbs are tested in two independent experiments. In the first experiment, with N bulbs, the exact lifetime $y_1, \ldots, y_N$ are recorded. In the second experiment, the experimenter enters the laboratory at some time $t > 0$, and all she registers is that some of the M litebulbs are still burning, while the others have expired. Thus, the results from the second experiment are right- or left-censored, and the available data are indicators $E_1, \ldots, E_M$”

$$E_i = \begin{cases} 1, & \text{if the bulb } i \text{ is still burning,} \\ 0, & \text{if light is out.} \end{cases}$$

[FZ] First Exercise.

The observed data from both the experiments combined denote

$$\mathbf{y} = (y_1, \ldots, y_N, E_1, \ldots, E_M)$$

and the unobserved data is

$$X = (X_1, \ldots, X_M).$$

The complete log-likelihood is

$$\begin{aligned} \ell^c(\theta; \mathbf{y}, X) &= \log\left(\prod_{i=1}^{N} \frac{\exp(-y_i/\theta)}{\theta} \prod_{i=1}^{M} \frac{\exp(-X_i/\theta)}{\theta}\right) \\ &= -N\big(\ln\theta + \bar{y}/\theta\big) - \sum_{i=1}^{M}\big(\ln\theta + X_i/\theta\big), \end{aligned}$$

which is linear in the unobserved X_i. But

$$\mathbb{E}(X_i \mid \mathbf{y}) = \mathbb{E}(X_i \mid E_i) = \begin{cases} t + \theta, & \text{if } E_i = 1, \\ \theta - \frac{t\exp(-t/\theta)}{1-\exp(-t/\theta)}, & \text{if } E_i = 0. \end{cases}$$

[FZ] First Exercise.

The E-step consists of replacing X_i by its expected value $\mathbb{E}(X_i \mid \mathbf{y})$ using the current value θ_t. Denote $Z = \sum_{i=1}^{M} Z_i$. Thus

$$
\begin{aligned}
Q(\theta \mid \theta_t) &= \mathbb{E}\ell^c(\theta; \mathbf{y}, X) = -N\Big(\ln\theta + \bar{y}/\theta\Big) - \sum_{i=1}^{M}\Big(\ln\theta + \mathbb{E}(X_i \mid E_i)/\theta\Big) \\
&= -(N+M)\ln\theta - \frac{1}{\theta}\left(N\bar{y} + Z(t+\theta_t) + (M-Z)\left(\theta_t - \frac{t\exp(-t/\theta_t)}{1-\exp(-t/\theta_t)}\right)\right).
\end{aligned}
$$

The M-step yields

$$
\begin{aligned}
\theta_{t+1} &= F(\theta_t) = \arg\max_{\theta} Q(\theta \mid \theta_t) \\
&= \frac{1}{N+M}\left(N\bar{y} + Z(t+\theta_t) + (M-Z)\left(\theta_t - \frac{t\exp(-t/\theta_t)}{1-\exp(-t/\theta_t)}\right)\right)
\end{aligned}
$$

[FZ] First Exercise.

"The self-consistency equation $\theta = F(\theta)$ has no explicit solution unless $Z = M$ (i.e., all litebulds in the second experiment are still on at time t); in this case, we obtain the well-known solution

$$\widehat{\theta} = \frac{N\bar{y} + Mt}{N}.$$

"

[FZ] Second Exercise.

"Contrary to litebulbs, lifetime of havybulbs follow a uniform distribution in the interval $(0, \theta]$, where θ is unknown. Suppose the same experiments are performed as in the first exercise, and again the second experimenter registers only that Z out of M havybulbs are still burning at time t, while $M - Z$ have expired.

... We know that for (hypothetical) complete data, the MLE would be max$\{Y_{max}, X_{max}\}$, where Y_{max} is the largest of the observed lifetimes, and X_{max} is the largest of the unobserved lifetimes."

[FZ] Second Exercise.

"Assume for simplicity that $Z \geq 1$, so that we are sure that $\theta \geq t$. Then

$$\mathbb{E}(X_i \mid E_i) = \begin{cases} \frac{1}{2}(t+\theta), & \text{if } E_i = 1, \\ \frac{1}{2}t, & \text{if } E_i = 0, \end{cases}$$

Thus, following the "rule" (substitute the X_i by its expectation in maximum likelihood estimator) we obtain

$$\theta_{t+1} = F(\theta_t) \equiv \max\left\{Y_{max}, \frac{1}{2}(t+\theta_t)\right\}.$$

"Starting with some $\theta_0 > 0$, iterations will converge to the solution $\widehat{\theta} = \max\{Y_{max}, t\}$, and this conclusion may be obtained easily by noticing that the self-consistency equation $\theta = F(\theta)$ is solved by $\widehat{\theta}$.

The main advantage of this solution is its simplicity. Its main disadvantage is that it is wrong."

[FZ] Second Exercise.

The joint likelihood function for the observed data is

$$L(\theta) = \theta^{-N}\mathbb{1}_{[Y_{max},\infty)}(\theta)\Big(\frac{t}{\max(t,\theta)}\Big)^{M-Z}\Big(1-\frac{t}{\max(t,\theta)}\Big)^{Z}.$$

Note that if $Z = 0$, then

$$L(\theta) = \theta^{-N}\mathbb{1}_{[Y_{max},\infty)}(\theta)\Big(\frac{t}{\max(t,\theta)}\Big)^{M},$$

“which is decreasing for $\theta \geq Y_{max}$, and therefore the maximum likelihood estimator is $\widehat{\theta} = Y_{max}$.

[FZ] Second Exercise. The joint likelihood function for the observed data is

$$L(\theta) = \theta^{-N} \mathbb{1}_{[Y_{max},\infty)}(\theta) \left(\frac{t}{\max(t,\theta)}\right)^{M-Z} \left(1 - \frac{t}{\max(t,\theta)}\right)^{Z}.$$

Note that if $Z \geq 1$, then $\theta \geq t$, and

$$\begin{aligned} L(\theta) &= \theta^{-N} \mathbb{1}_{[Y_{max},\infty)}(\theta) \left(\frac{t}{\theta}\right)^{M-Z} \left(1 - \frac{t}{\theta}\right)^{Z} \\ &= t^{M-Z} \mathbb{1}_{[Y_{max},\infty)}(\theta) \theta^{-(N+M)} (\theta - t)^{Z}. \end{aligned}$$

For $\theta \geq t$ the function $\theta^{-(N+M)}(\theta - t)^Z$ has a unique maximum in $\bar{\theta} = \frac{N+M}{N+M-Z} t$ and is monotonically decreasing for $\theta \geq \bar{\theta}$. Thus summarizing the results the likelihood function estimator is

$$\widehat{\theta} = \begin{cases} \bar{\theta}, & \text{if } \bar{\theta} > Y_{max} \text{ and } Z \geq 1, \\ Y_{max}, & \text{otherwise.} \end{cases} \qquad (\widehat{\theta} = \max\{Y_{max}, t\})$$

[FZ] Second Exercise.

"Why is the solution given by the EM algorithm wrong? The answer is simple: the EM algorithm in not applicable because the log-likelihood function does not exist for all $\theta > 0$, which means that its expected value is not defined."

[FZ] Second Exercise.

"Indeed, assume that one heavybulb has survived time t, and let X_m be its (unobserved) lifetime. The unconditioned distribution of X_m is $U[0, \theta]$. In E-step we need to find $Q(\theta \mid \theta_t)$. The conditional expectation of X_m is calculated conditioning on the event $X_m > t$ and using θ_t as a parameter, thus $X_m|Y$ has uniform $U[t, \theta_t]$ distribution. Now, for all $\theta < \theta_t$ the unconditioned density of X_m

$$f_\theta(x) = \begin{cases} 1/\theta, & \text{if } 0 \le x \le \theta, \\ 0, & \text{elsewhere.} \end{cases}$$

takes value 0 with positive probability, and hence $Q(\theta \mid \theta_t)$ does not exist for $\theta < \theta_t$. This could be seen from the observed data likelihood function, but in the rush of applying the EM algorithm, it is easy to skip this check."

[H] About Second Exercise of [FZ].

Let the EM algorithm start at some $\theta_0 > \max\{Y_{max}, t\}$. In [H] it is shown that "the EM algorithm in this example converges to θ_0 – in other words, it never goes anywhere once initialized!"

$$Q(\theta \mid \theta_0) = \begin{cases} -(N+M)\log\theta, & \text{if } \theta \geq \theta_0, \\ -\infty, & \text{if } 0 < \theta < \theta_0. \end{cases}$$

"Since $Q(\theta \mid \theta_0)$ is strictly decreasing on $[\theta_0, \infty)$ and strictly less than $Q(\theta_0 \mid \theta_0)$ on $(0, \theta_0)$, setting θ_1 equal to the maximizer of $Q(\theta \mid \theta_0)$ gives $\theta_1 = \theta_0$. By induction, this EM algorithm is forever stuck at the initial value."

References.

[DLR] Dempster, A.P., Laird, N.M., and Rubin, D.B. *Maximum likelihood from incomplete data via the EM algorithm*, J.Roy. Statist. Soc. Ser. B, **39**, 1-38, 1977.

[RC] Cristian P. Robert and George Casella. *Introducing Monte Carlo Methods with R*. Series "Use R!". Springer

[BH] Booth J.G. and Hobert J.P. *Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm*. J.R. Statist. Soc. B. **61**, Part1, pp. 265-285, 1999.

[FZ] Flury, Bernard and Zoppé, Alice. *Exercises in EM*. The American Statistician, Vol. 54, No.3, August 2000.

[H] Hunter, David R. *On the Geometry of EM algorithms.* Technical Report 0303, Penn State University, 2003.