# 8 Notes on EM Algorithm

## 8.1 EM Algorithm for Multinomial & Mixture Data

**General Example 1.** Suppose that for fixed integers $1 \leq K < C$, cell-counts $\mathbf{X} = (X_1, \ldots, X_K)$ are observed, and cell-counts $\mathbf{Y} = (Y_{K+1}, \ldots, Y_C)$ *cannot* be observed, where

$$\Big( X_1, \, \ldots, \, X_K, \, Y_{K+1}, \ldots, \, Y_C \Big) \; \sim \; \textit{Multinomial}(n, \, p_j(\vartheta), \; j = 1, \ldots, C)$$

Here $\vartheta$ is an unknown parameter of dimension $d \leq K$, and the functions $p_j(\vartheta)$ which share $\vartheta$ as a parameter are sufficiently smooth. Also denote

$$X_{K+1} \; = \; n - X_1 - \cdots - X_K \; = \; \sum_{j=K+1}^{C} Y_j$$

For notational convenience, define

$$q_K(\vartheta) \; = \; 1 - p_1(\vartheta) - \cdots - p_K(\vartheta)$$

In this setting, we express the conditional joint density of $\mathbf{Y}$ given $\mathbf{X}$ by

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X}, \vartheta) \; = \; \exp( \sum_{j=K+1}^{C} y_j \, \log \big( \frac{p_j(\vartheta)}{q_K(\vartheta)} \big) ) \; \cdot \; \binom{X_{K+1}}{Y_{K+1}, \ldots, Y_C}$$

It follows that the *E-step* of the EM algorithm replaces $E_{\vartheta_1} \Big( \log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X}, \, \vartheta) \Big)$ by

$$X_{K+1} \sum_{j=K+1}^{C} \frac{p_j(\vartheta_1)}{q_K(\vartheta_1)} \, \log \big( \frac{p_j(\vartheta)}{q_K(\vartheta)} \big) \; + \; \log \binom{X_{K+1}}{Y_{K+1}, \ldots, Y_C}$$

or equivalently, replaces $Y_j$ by $X_{K+1} \cdot p_j(\vartheta_1)/q_K(\vartheta_1)$ for $j = K+1, \ldots, C$.

To confirm that the definition of log-likelihood and conditional log-likelihood terms as above, without multinomial coefficients, is legitimate, we observe that the property needed in the proof of log-likelihood improvement for EM iterations holds, that is,

$$E_\vartheta \Big( \log L_{\mathbf{Y}|X}(n_{C+1}, \ldots, n_K \,|\, \mathbf{X}, \vartheta) - \log L_{\mathbf{Y}|X}(n_{C+1}, \ldots, n_K \,|\, \mathbf{X}, \vartheta_1) \Big) \geq 0$$

or equivalently, for all $\vartheta$, $\vartheta_1$,

$$\sum_{j=C+1}^{K} \frac{p_j(\vartheta)}{q(\vartheta)} \log \frac{p_j(\vartheta)\, q(\vartheta_1)}{q(\vartheta)\, p_j(\vartheta_1)} \geq 0$$

But this is a standard, discrete version of the famous 'Information Inequality' proved more generally in the form $\int f(x) \log(f(x)/g(x))\, d\nu(x) \geq 0$ for probability densities with respect to a measure $\nu$, using Jensen's Inequality.

In the multinomial setting, we express the conditional likelihood for $\mathbf{Y}$ given $\mathbf{X}$ by

$$L_{\mathbf{Y}|\mathbf{X}}(n_{C+1}, \ldots, n_K \,|\, \mathbf{X}, \vartheta) \;=\; \exp( \sum_{j=C+1}^{K} n_j \log \big( \frac{p_j(\vartheta)}{q(\vartheta)} \big) )$$

The *E-step* of the EM algorithm replaces this conditional log-likelihood, when the current parameter-iterate is $\vartheta_1$, by

$$E_{\vartheta_1}\Big( \log L_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X},\, \vartheta) \;\Big|\; \mathbf{X} \Big) \quad = \quad n^* \sum_{j=C+1}^{K} \frac{p_j(\vartheta_1)}{q(\vartheta_1)} \log \Big( \frac{p_j(\vartheta)}{q(\vartheta)} \Big)$$

This expression is also equal to $\log L_{\mathbf{Y}|\mathbf{X}}(n^*_{C+1}, \ldots, n^*_K \,|\, \mathbf{X}, \vartheta)$, where $n^*_j = E_{\vartheta_1}(n^*_j \,|\, \mathbf{X}) \;=\; n^* \cdot p_j(\vartheta_1)/q(\vartheta_1)$ for $j = C+1, \ldots, K$.

Thus we have the following comparison between maximization approaches. First, the complete-data likelihood to maximize, if $\mathbf{Y}$ could also be observed, would be

$$\sum_{j=1}^{K} X_j \log p_j(\vartheta) \;+\; X_{K+1} \log q_K(\vartheta) \;+\; \sum_{j=K+1}^{C} Y_j \log \frac{p_j(\vartheta)}{q_K(\vartheta)}$$

while the crude marginal-observed-data likelihood to maximize is

$$\sum_{j=1}^{K} X_j \log p_j(\vartheta) \;+\; X_{K+1} \log q_K(\vartheta)$$

On the other hand, the *M-step* of the EM algorithm, after replacement of the unobservable $Y_j$ values in the complete-data likelihood by their *E-step* imputed values, is

$$\sum_{j=1}^{K} X_j \log p_j(\vartheta) \;+\; X_{K+1} \log q_K(\vartheta) \;+\; X_{K+1} \sum_{j=K+1}^{C} \frac{p_j(\vartheta_1)}{q_K(\vartheta_1)} \log \frac{p_j(\vartheta)}{q_K(\vartheta)}$$

$$= \sum_{j=1}^{K} X_j \, \log p_j(\vartheta) \; + \; X_{K+1} \sum_{j=K+1}^{C} \frac{p_j(\vartheta_1)}{q_K(\vartheta_1)} \, \log p_j(\vartheta)$$

Note that the M-step involves a step of maximizing the complete-data likelihood using imputed data for the $Y_j$'s, which will be very easy in some problems.

*A key aspect of the usefulness of the EM algorithm in multinomial missing data problems is that no sums of terms* $p_j(\vartheta)$ *appear inside the logarithms arising in the maximization-step. Especially in so-called log-linear contingency-table models with some missing cell-counts, where the* $p_j(\vartheta)$ *have some multiplicative structure, this is very useful !*

*SPECIAL EXAMPLE FROM THE ORIGINAL EM PAPER*

This example fits into the structure of the general multinomial example, with scalar unknown parameter $\vartheta = \pi$, $K = 3$, $C = 5$, and

$$p_1(\pi) \; = \; p_2(\pi) \; = \; \frac{1-\pi}{4} \, , \quad p_3(\pi) \; = \; p_4(\pi) \; = \; \frac{\pi}{4} \, , \quad p_5(\pi) \; = \; \frac{1}{2}$$

The cell-counts given as data in Dempster, Laird & Rubin (1978) are:

$(X_1, X_2, X_3, X_4) = (18, 20, 34, 125)$. Appealing to the formulas above, we find that the complete-data M-step involves maximizing $\sum_{j=1}^{3} X_j \log p_j(\pi) + \sum_{j=4}^{5} Y_j \, \log p_j(\pi)$. In this particular problem , we are equivalently maximizing $(X_3 + Y_4) \, \log(\pi/4) \, + \, (X_1 + X_2) \, \log((1-\pi)/4)$, which leads to

$$\hat{\pi} \; = \; (X_3 + Y_4)/(n - Y_5)$$

Substituting the E-step imputed valued for the $Y_j$ gives the EM iteration explicitly, starting from initial guess $\pi_1$, as:

$$\pi_2 \; = \; \Big(X_3 + \, X_4 \cdot \frac{\pi_1/4}{1/2 + \pi_1/4}\Big) \Big/ \Big(n - X_4 \cdot \frac{1/2}{1/2 + \pi_1/4}\Big)$$

$$= \; \frac{34 + 125 \cdot \frac{\pi_1}{2+\pi_1}}{197 - 125 \cdot \frac{2}{2+\pi_1}} \; = \; \frac{68 \; + \; 159 \, \pi_1}{144 + 197 \, \pi_1}$$

In this little example, EM iterates the mapping $h(\pi) \equiv (68 + 159\pi)/(144 + 197\pi)$ to find the fixed-point. (The unique fixed-point $\pi = 0.6268$ solves $h(\pi) = \pi$, which is a quadratic equation.) The Quasi-Newton optimization of the marginal likelihood is messier but, using a modern computer, quicker and more reliable.

```
> optimize(function(x) 38*log(1-x)+34*log(x)+125*log(x+2),
  c(.01,.99), max=T)$max
[1] 0.6268036
> h <- function(x) (159 * x + 68)/(197 * x + 144)
  x<- .5; for (i in 1:6) {x <- h(x)  cat(round(x,5)," \n")}
0.60825
0.62432
0.62649
0.62678
0.62682    ### converged to 5 places after 5 iterations
```

**General Example 2.** Consider 'mixture' data $X_i$ which are *iid* continuously distributed *rv*'s with density

$$f_X(x) \;=\; pe^{-x} \,+\, \lambda\,(1-p)\,e^{-\lambda x} \quad , \qquad x > 0$$

where $\vartheta = (p, \lambda) \in (0,1) \times [0, \infty)$ is the unknown parameter. These r.v.'s are of *mixture* type because they have the same density as random variables

$$X_i \;=\; \epsilon_i\, U_i \,+\, (1-\epsilon_i)\, V_i \qquad U_i \sim \mathit{Expon}(1)\;, \quad V_i \sim \mathit{Expon}(\lambda)$$

where $\epsilon_i \sim \mathit{Binom}(1,p)$ is independent of $(U_i, V_i)$. The marginal density for the observed variables is $f_X$, but the problem would be much simpler to analyze with the 'complete' data $(X_i,\, \epsilon_i),\ i = 1, \ldots, n$. Now the *E-step* of the EM algorithm based on observing only $\mathbf{X} = (X_i,\ i = 1, \ldots, n)$ consists of calculating

$$E_{\vartheta_1}(\epsilon \,|\, X) \;=\; \frac{p_1\, e^{-X}}{p_1\, e^{-X} \,+\, \lambda_1\,(1-p_1)\, e^{-\lambda_1 X}} \;=\; \epsilon^*(X, \vartheta_1) \;=\; \epsilon^*$$

and then substituting to obtain

$$E_{\vartheta_1} \log p_{\epsilon|X}(\epsilon \,|\, X, \vartheta)) \;=\; \epsilon^* \log\Big(\frac{p\, e^{-X}}{p\, e^{-X} \,+\, \lambda\,(1-p)\, e^{-\lambda X}}\Big)$$

$$+\; (1-\epsilon^*)\, \log\Big(\frac{\lambda\,(1-p)\, e^{-\lambda X}}{p\, e^{-X} \,+\, \lambda\,(1-p)\, e^{-\lambda X}}\Big)$$

As a result, starting from initial guess $\vartheta_1 = (\lambda_1, p_1)$, the *M-step* of the EM algorithm is to maximize the 'complete-data log-likelihood' for the data

$(X_i,\, \epsilon^*(X_i, \vartheta_1),\ i = 1, \ldots, n)$, which is given simply in terms of

$$m^* \;=\; \sum_{i=1}^n \epsilon^*(X_i, \vartheta_1) \quad , \qquad \overline{U} \;=\; (m^*)^{-1} \sum_{i=1}^n X_i\, \epsilon^*(X_i, \vartheta_1)$$

and

$$\overline{V} \;=\; (n - m^*)^{-1} \sum_{i=1}^n X_i\, (1 - \epsilon^*(X_i, \vartheta_1))$$

as

$$m^*\, (\log p \,-\, \overline{U}) \,+\, (n - m^*)\, (\log(\lambda(1-p)) \,-\, \lambda \overline{V})$$

Thus the *M-step* is given in closed form by maximizing the last expression in $(\lambda,\, p)$ to obtain

$$p_2 \;=\; m^*/n \quad , \qquad \lambda_2 \;=\; 1/\overline{V}$$

In summary, the entire EM iteration-step in this example, starting from initial guess $\vartheta_1 \;=\; (\lambda_1,\, p_1)$, is given in closed form by:

$$p_2 \;=\; \frac{1}{n} \sum_{i=1}^n \frac{p_1\, e^{-X_i}}{p_1\, e^{-X_i} \,+\, \lambda_1\, (1 - p_1)\, e^{-\lambda_1 X_i}}$$

$$1/\lambda_2 \;=\; \frac{1}{n(1 - p_2)} \sum_{i=1}^n \frac{(1 - p_1)\lambda_1\, X_i\, e^{-\lambda_1 X_i}}{p_1\, e^{-X_i} \,+\, \lambda_1\, (1 - p_1)\, e^{-\lambda_1 X_i}}$$

We implement this, and evaluate the results in a little simulated dataset, as follows.

```
> EMiter
function(thet, Xvec)
{
## On input, thet is the vector consisting of old values of
##    p, lambda in General Example 2 of Notes, and Xvec is
##    the observed data vector. The output is the new theta.
  frac <- 1/(1 + (1/thet[1] - 1) * thet[2] * exp((
        1 - thet[2]) * Xvec))
  pnew <- mean(frac)
  lamnew <- (1 - pnew)/mean(Xvec * (1 - frac))
```

```
  list(thet = c(pnew, lamnew), logL = sum(log(pnew * exp(
     - Xvec) + (1 - pnew) * lamnew * exp( - lamnew * Xvec))))
}


> epsv <- rbinom(10000, 1, .6)
  Xv <- rexp(10000)/exp(.3*(1-epsv))
> round(c(mean(epsv), .4/exp(.3)+.6, mean(Xv)),5)
[1] 0.60050 0.89633 0.89459

> theta <- c(.5,1.5)
## Initial log-likelihood
> sum(log(.5 * exp( - Xv) + .5*1.5*exp(-1.5*Xv))) ## = -8908.9
## Log-likelihood at true values:
> sum(log(.6 * exp( - Xv) + .4*exp(.3-exp(.3)*Xv)))##  -8883.5

> unlist(EMiter(theta,Xv))   ## values after one EM iteration
    thet1    thet2      logL
 0.5069983 1.413503 -8893.04

> for(i in 1:100) {
       tmpitr <- EMiter(theta,Xv)
       theta <<- tmpitr$thet
       if(i %% 5 ==0) cat(round(unlist(tmpitr),5),"\n") }
       0.51477 1.30256 -8883.91956
0.51623 1.28736 -8883.74519
0.51689 1.28652 -8883.74114
0.51749 1.28686 -8883.73824
0.51807 1.28728 -8883.73536
0.51866 1.28772 -8883.73248
0.51925 1.28815 -8883.72962
0.51983 1.28859 -8883.72677
0.52042 1.28903 -8883.72393
0.521 1.28947 -8883.72111
0.52158 1.2899 -8883.71829
0.52216 1.29034 -8883.71549
0.52274 1.29078 -8883.7127
0.52331 1.29122 -8883.70992
```

```
0.52389 1.29165 -8883.70716
0.52446 1.29209 -8883.7044
0.52504 1.29253 -8883.70166
0.52561 1.29297 -8883.69893
0.52618 1.29341 -8883.69622
0.52675 1.29385 -8883.69351

> for(i in 1:100) theta <- EMiter(theta,Xv)$thet
  unlist(EMiter(theta,Xv))
     thet1    thet2      logL
 0.5379237 1.302709 -8883.642
> for(i in 1:100) theta <- EMiter(theta,Xv)$thet
  unlist(EMiter(theta,Xv))
>    thet1    thet2      logL
 0.5483496 1.311421 -8883.596
## Convergence is painfully slow !!!

> nlminb(c(0.5, 1.5), function(x)  - sum(log(x[1] * exp( - Xv) +
     (1 - x[1]) * x[2] * exp( - x[2] * Xv))), lower
         = c(0.01, 0.1), upper = c(0.99, 10))[1:4]
$parameters:
[1] 0.6308609 1.3997779

$objective:
[1] 8883.386

$message:
[1] "RELATIVE FUNCTION CONVERGENCE"

$grad.norm:
[1] 0.001797887
```

Note the very slow convergence of the EM algorithm implemented and tested here. The maximized *logLik* must be larger than $-8883.5$, since that is the value at the true parameters ($p = .6$, $\lambda = e^{.3}$), but from the not-too-awful starting values $p_1 = .5$, $\lambda = 1.5$, it took more than 300 EM iterations to get there ! As can be seen from the final converged maximization via **nlminb**, the final maximized *logLik* is $-8883.39$.