Statistical Methods for Handling Incomplete Data
Chapter 3.3: EM algorithm

Jae-Kwang Kim

Motivation

- By Theorem 2.5, solving $S_{obs}(\eta) = 0$ is equivalent to solving $\bar{S}(\eta) = 0$.
- EM algorithm provides an alternative method of solving $\bar{S}(\eta) = 0$ by writing

$$\bar{S}(\eta) = E\left\{S_{com}(\eta) \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \eta\right\}$$

  and using the following iterative method:

$$\hat{\eta}^{(t+1)} \leftarrow \text{ solve } E\left\{S_{com}(\eta) \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \hat{\eta}^{(t)}\right\} = 0.$$

  - E-step: Compute the conditional expectation given the observed data evaluated at $\hat{\eta}^{(t)}$
  - M-step: Update the parameter by solving the above mean score equation.

Definition

Let $L_{com}(\eta)$ be the likelihood function of η based on the complete-sample observations. The EM algorithm is an iterative algorithm defined by the following E-step and M-steps:

- E-step: Compute

$$Q\left(\eta \mid \eta^{(t)}\right) = E\left\{\log L_{\rm com}(\eta) \mid \mathbf{y}_{\rm obs}, \boldsymbol{\delta}, \eta^{(t)}\right\},$$

  where $\eta^{(t)}$ be the current value of the parameter estimate of η.
- M-step: Find $\eta^{(t+1)}$ that maximizes $Q(\eta \mid \eta^{(t)})$ w.r.t. η.

Theorem 3.2 (Dempster et al., 1977)

Let $L_{obs}(\eta) = \int f(\mathbf{y}, \boldsymbol{\delta}; \eta)\, d\mathbf{y}_{\rm mis}$ be the observed likelihood of η. If $Q(\eta^{(t+1)} \mid \eta^{(t)}) \geq Q(\eta^{(t)} \mid \eta^{(t)})$, then $L_{obs}(\eta^{(t+1)}) \geq L_{obs}(\eta^{(t)})$.

By Theorem 3.2, the sequence $\{L_{\rm obs}(\eta^{(t)})\}$ is monotone increasing and it is bounded above if the MLE exists. Thus, the sequence of $L_{\rm obs}(\eta^{(t)})$ converges to some value L^*. In most cases, L^* is a stationary value in the sense that $L^* = L_{\rm obs}(\eta^*)$ for some η^* at which $\partial L_{\rm obs}(\eta)/\partial\eta = 0$. Under fairly weak conditions, such as $Q(\eta \mid \gamma)$ satisfies

$$\partial Q(\eta \mid \gamma)/\partial\eta \text{ is continuous in } \eta \text{ and } \gamma,$$

the EM sequence $\{\eta^{(t)}\}$ converges to a stationary point η^*. (Wu, 1983)

Proof of Theorem 3.2

EM for regression with missing Y

- The parameter of interest is θ in $f(y \mid x; \theta)$.
- Under complete response, the log-likelihood function for θ is given by

$$\ell_{com}(\theta) = \sum_{i=1}^{n} \log f(y_i \mid x_i; \theta)$$

- Let $\delta_i \overset{iid}{\sim}$ Beroulli$\{\pi(x, y)\}$.
- Assume that y_i is observed if and only if $\delta_i = 1$

- E-step: Using the current parameter $\theta^{(t)}$, compute

$$\begin{aligned} & Q(\theta \mid \theta^{(t)}) \\ \equiv \quad & E\left\{\ell_{com}(\theta) \mid \text{data}, \theta^{(t)}\right\} \\ = \quad & \sum_{i=1}^{n} \delta_i \log f(y_i \mid x_i; \theta) + \sum_{i=1}^{n}(1-\delta_i) E\left\{\log f(Y \mid x_i; \theta) \mid x_i, \delta_i = 0; \theta^{(t)}\right\} \\ := \quad & Q_1(\theta \mid \theta^{(t)}) + Q_2(\theta \mid \theta^{(t)}) \end{aligned}$$

where the conditional distribution is with respect to the prediction model

$$f(y \mid x, \delta = 0; \theta^{(t)}) = \frac{f(y \mid x; \theta^{(t)})\{1 - \pi(x, y)\}}{\int f(y \mid x; \theta^{(t)})\{1 - \pi(x, y)\} dy}. \tag{1}$$

- M-step: Update the parameter by finding the maximizer of $Q(\theta \mid \theta^{(t)})$.

Remark

- Under MAR, we have $\pi(x, y) = \pi(x)$. In this case, the prediction model in (1) is changed to

$$f(y \mid x, \delta = 0; \theta^{(t)}) \quad = \quad f(y \mid x; \theta^{(t)}).$$

- Note that

$$E\left\{\log f(Y \mid x; \theta) \mid x, \delta = 0; \theta^{(t)}\right\} = E\left\{\log f(Y \mid x; \theta) \mid x; \theta^{(t)}\right\}$$

is maximized at $\theta = \theta^{(t)}$ (by Lemma 2.1), which means that the second term of $Q(\theta \mid \theta^{(t)})$ does not contribute to the parameter estimation of θ.

- Therefore, we have only to use the cases with $\delta_i = 1$ to obtain the maximum likelihood estimator

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} \delta_i \log f(y_i \mid x_i; \theta).$$

Categorical Missing data

If $\mathbf{y}$ is a categorical variable that takes values in set S_y, then the E-step can be easily computed by a weighted summation

$$E\left\{\log f\left(\mathbf{y}, \boldsymbol{\delta}; \eta\right) \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}\right\} = \sum_{\mathbf{y} \in S_y} P(\mathbf{y} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}) \log f\left(\mathbf{y}, \boldsymbol{\delta}; \eta\right) \qquad (2)$$

where the summation is over all possible values of $\mathbf{y}$ and $P\left(\mathbf{y} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}, \eta^{(t)}\right)$ is the conditional probability of taking $\mathbf{y}$ given $\mathbf{y}_{\mathrm{obs}}$ and $\boldsymbol{\delta}$ evaluated at $\eta^{(t)}$. The conditional probability $P\left(\mathbf{y} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\delta}; \eta^{(t)}\right)$ can be treated as the weight assigned for the categorical variable $\mathbf{y}$. That is, if $S(\eta) = \sum_{i=1}^{n} S(\eta; \mathbf{y}_i, \delta_i)$ is the score function for η, then the EM algorithm using (2) can be obtained by solving

$$\sum_{i=1}^{n} \sum_{\mathbf{y} \in S_y} P\left(\mathbf{y}_i = \mathbf{y} \mid \mathbf{y}_{i,\mathrm{obs}}, \boldsymbol{\delta}_i, \eta^{(t)}\right) S(\eta; \mathbf{y}, \delta_i) = 0$$

for η to get $\eta^{(t+1)}$. Ibrahim (1990) called this approach *EM by weighting*.

- E-step:

$$\bar{S}_1\left(\beta \mid \beta^{(t)}, \phi^{(t)}\right) = \sum_{\delta_i=1}\{y_i - p_i(\beta)\}\mathbf{x}_i + \sum_{\delta_i=0}\sum_{j=0}^{1} w_{ij(t)}\{j - p_i(\beta)\}\mathbf{x}_i,$$

where

$$\begin{aligned} w_{ij(t)} &= Pr(Y_i = j \mid \mathbf{x}_i, \delta_i = 0; \beta^{(t)}, \phi^{(t)}) \\ &= \frac{Pr(Y_i = j \mid \mathbf{x}_i; \beta^{(t)})Pr(\delta_i = 0 \mid \mathbf{x}_i, j; \phi^{(t)})}{\sum_{y=0}^{1} Pr(Y_i = y \mid \mathbf{x}_i; \beta^{(t)})Pr(\delta_i = 0 \mid \mathbf{x}_i, y; \phi^{(t)})} \end{aligned}$$

and

$$\begin{aligned} \bar{S}_2\left(\phi \mid \beta^{(t)}, \phi^{(t)}\right) &= \sum_{\delta_i=1}\{\delta_i - \pi(\mathbf{x}_i, y_i; \phi)\}(\mathbf{x}_i', y_i)' \\ &\quad + \sum_{\delta_i=0}\sum_{j=0}^{1} w_{ij(t)}\{\delta_i - \pi_i(\mathbf{x}_i, j; \phi)\}(\mathbf{x}_i', j)'. \end{aligned}$$

- M-step:
  The parameter estimates are updated by solving

$$\left[\bar{S}_1\left(\beta \mid \beta^{(t)}, \phi^{(t)}\right), \bar{S}_2\left(\phi \mid \beta^{(t)}, \phi^{(t)}\right)\right] = (0, 0)$$

  for β and ϕ.
- Thus, the conditional expectation in the E-step can be computed using the weighted mean with weights $w_{ij(t)}$.
- Observed information matrix can also be obtained by the Louis formula (in Theorem 2.7) using the weighted mean in the E-step.

EM in the exponential family

Under MAR and for the exponential family of the distribution of the form

$$f(\mathbf{y};\theta) = b(\mathbf{y})\, exp\left\{\theta'\mathbf{T}(\mathbf{y}) - A(\theta)\right\}.$$

Under MAR, the E-step of the EM algorithm is

$$Q\left(\theta \mid \theta^{(t)}\right) = \text{ constant } + \theta' E\left\{\mathbf{T}(\mathbf{y}) \mid \mathbf{y}_{\text{obs}}, \theta^{(t)}\right\} - A(\theta) \tag{3}$$

and the M-step is

$$\frac{\partial}{\partial\theta} Q\left(\theta \mid \theta^{(t)}\right) = 0 \quad \Longleftrightarrow \quad E\left\{\mathbf{T}(\mathbf{y}) \mid \mathbf{y}_{\text{obs}}, \theta^{(t)}\right\} = \frac{\partial}{\partial\theta} A(\theta).$$

Because $\int f(\mathbf{y};\theta)\, d\mathbf{y} = 1$, we have

$$\frac{\partial}{\partial\theta} A(\theta) = E\left\{\mathbf{T}(\mathbf{y})\,;\theta\right\}.$$

Therefore, the M-step reduces to finding $\theta^{(t+1)}$ as a solution to

$$E\left\{\mathbf{T}(\mathbf{y}) \mid \mathbf{y}_{\text{obs}}, \theta^{(t)}\right\} = E\left\{\mathbf{T}(\mathbf{y})\,;\theta\right\}. \tag{4}$$

Graphical Illustration

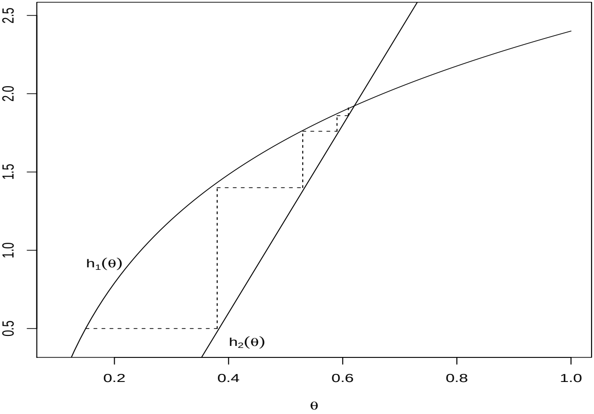

Figure: Illustration of EM algorithm for exponential family ($h_1(\theta) = E\{\mathbf{T}(\mathbf{y}) \mid \mathbf{y}_{\text{obs}}, \theta\}$, $h_2(\theta) = E\{\mathbf{T}(\mathbf{y}) \mid \theta\}$)

Example 3.9 (Bivariate Normal distribution)

- Model

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}\right]$$

- Sufficient statistics

$$S = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i y_i, \sum_{i=1}^{n} y_i^2\right)$$

- The EM algorithm reduces to solving

$$\sum_{i=1}^{n} E\left\{\left(x_i, y_i, x_i^2, x_i y_i, y_i^2\right) \mid \delta_i^{(x)}, \delta_i^{(y)}, \delta_i^{(x)} x_i, \delta_i^{(y)} y_i; \theta^{(t)}\right\}$$

$$= \sum_{i=1}^{n} E\left\{\left(x_i, y_i, x_i^2, x_i y_i, y_i^2\right); \theta\right\}$$

for $\theta = (\mu_x, \mu_y, \sigma_{xx}, \sigma_{xy}, \sigma_{yy})'$. Under MAR, the above conditional expectation can be obtained using the usual conditional expectation under normality.

Example 3.6

Table: **A 2 $\times$ 2 table with supplemental margins for both variables**

Set	y_1	y_2	Count
H	1	1	100
	1	2	50
	2	1	75
	2	2	75
K	1		30
	2		60
L		1	28
		2	60

Example 3.6 (Cont'd)

- The parameters of interest are $\pi_{ij} = P\left(Y_1 = i,\ Y_2 = j\right),\ i = 1, 2, j = 1, 2.$
- The sufficient statistics for the parameters are $n_{ij}, i = 1, 2; j = 1, 2$, where n_{ij} is the sample size for the set with $Y_1 = i$ and $Y_2 = j$.
- The E-step computes the conditional expectation of the sufficient statistics. This gives

$$n_{ij}^{(t)} = E\left(n_{ij} \mid \text{data}, \pi_{ij}^{(t)}\right) = n_{ij,H} + n_{i+,K}\frac{\pi_{ij}^{(t)}}{\pi_{i+}^{(t)}} + n_{+j,L}\frac{\pi_{ij}^{(t)}}{\pi_{+j}^{(t)}},$$

  for $i = 1, 2; j = 1, 2.$
- In the M-step, the parameters are updated by $\pi_{ij}^{(t+1)} = n_{ij}^{(t)}/n$.

R program for EM algorithm in Example 3.6

```
> setH=matrix(c(100,75,50,75),2,2)
> setK=c(30,60)
> setL=c(28,60)
> th=prop.table(setH) #initial estimates of pi from setH
> round(th,3)
      [,1]  [,2]
[1,] 0.333 0.167
[2,] 0.250 0.250
> nij=matrix(nrow=2,ncol=2)
> repeat{
+ th0=th
+ #E-step
+ for(i in 1:2){
+ for(j in 1:2){
+ nij[i,j]=setH[i,j]+setK[i]*th[i,j]/sum(th[i,])+setL[j]*th[i,j]/sum(th[,j])
+ }}
+ #M-step
+ th=nij/n
+ dif=sum((th0-th)^2)
+ if(dif<1e-8) break}
> round(th,3)
      [,1]  [,2]
[1,] 0.279 0.174
[2,] 0.239 0.308
```

Example 3.12

- Model: $x_i = \mu + \sigma e_i$ with $e_i \sim t(\nu)$, ν: known.
- Missing data setup:

$$x_i \mid w_i \sim N\left(\mu, \sigma^2 / w_i\right), \quad w_i \sim \chi^2_\nu / \nu.$$

- (x_i, w_i): complete data
- x_i always observed, w_i always missing
- Parameter: $\theta = (\mu, \sigma)$

Example 3.12 (Cont'd)

E-step: Find the conditional distribution of w_i given x_i. By Bayes theorem,

$$
\begin{aligned}
f(w_i \mid x_i) &\propto f(w_i) f(x_i \mid w_i) \\
&\propto (w_i \nu)^{\frac{\nu}{2}-1} \exp\left(-\frac{w_i \nu}{2}\right) \times \left(\sigma^2 / w_i\right)^{-1/2} \exp\left\{-\frac{w_i}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\} \\
&\sim \text{Gamma}\left[\frac{\nu+1}{2}, 2\left\{\nu + \left(\frac{x_i-\mu}{\sigma}\right)^2\right\}^{-1}\right].
\end{aligned}
$$

Thus, the E-step of EM algorithm can be written as

$$
E(w_i \mid x_i, \theta^{(t)}) = \frac{\nu+1}{\nu + \left(d_i^{(t)}\right)^2},
$$

where $d_i^{(t)} = (x_i - \mu^{(t)})/\sigma^{(t)}$.

Example 3.12 (Cont'd)

M-step:

$$
\begin{aligned}
\mu^{(t+1)} &= \frac{\sum_{i=1}^{n} w_i^{(t)} x_i}{\sum_{i=1}^{n} w_i^{(t)}} \\
\sigma^{2(t+1)} &= \frac{1}{n} \sum_{i=1}^{n} w_i^{(t)} \left(x_i - \mu^{(t+1)} \right)^2
\end{aligned}
$$

where $w_i^{(t)} = E(w_i \mid x_i, \theta^{(t)})$.

Remark

- The EM algorithm in Example 3.12 can be further extended to the problem of robust regression, where the error distribution in the regression model is assumed to follow from a t-distribution. Suppose that the model for robust regression can be written as

$$y_i = \beta_0 + \beta_1 x_i + \sigma e_i$$

  where $e_i \sim t(\nu)$ with a known ν.
- Similarly to Example 3.12, we can write

$$y_i \mid (x_i, w_i) \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2 / w_i\right), \quad w_i \sim \chi^2_\nu / \nu.$$

  Here, (x_i, y_i) are always observed, and w_i are always missing.
- ν plays the role of tuning parameter. How to determine ν in practice?

REFERENCES

Dempster, A. P., N. M. Laird and D. B. Rubin (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B* **39**, 1–37.

Ibrahim, J. G. (1990), 'Incomplete data in generalized linear models', *Journal of the American Statistical Association* **85**, 765–769.

Wu, C. F. J. (1983), 'On the convergence properties of the EM algorithm', *The Annals of Statistics* **11**, 95–103.