# An Introduction of EM Algorithm

Yi-Hau Chen
Department of Mathematics
National Taiwan Normal University

# OUTLINE

1. Introduction
2. EM in exponential family
3. EM in general forms
4. General properties
5. Applications of EM algorithms to incomplete data analysis
6. References:
   Dempster, Laird and Rubin (JRSSB, pp1-37, 1977). Maximum Likelihood from Incomplete Data via the EM Algorithm.

# 1. Introduction

- EM algorithm:
  a general approach to iterative computation of MLEs when the observations can be viewed as incomplete data; each iteration of the algorithm consists of an expectation step followed by a maximization step. (E-step and M-step)

- incomplete data:
  $y$: observed data in the sample space $\mathcal{Y}$
  $x$: complete data in the sample space $\mathcal{X}$
  (including what are traditionally called parameters)
  There exists a many-one mapping: $x \to y(x)$ from $\mathcal{X}$ to $\mathcal{Y}$; $x$ is known only to lie in $\mathcal{X}(y)$.

- complete-data specification: $f(x|\phi)$ ($\phi$: parameters)

- incomplete data specification:

$$g(y|\phi) = \int_{\mathcal{X}(y)} f(x|\phi)dx$$

EM is directed at finding a value of $\phi$ maximizing $g(y|\phi)$ given an observed $y$, but it does so by making essential use of $f(x|\phi)$.

- An Illustrative Example
  (Rao, 1965, pp. 368-369)
  197 animals are distributed multinomially into four categories, the observed data

  $$\mathrm{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34).$$

  A genetic model specifies cell probabilities

  $$(1/2 + \phi/4, (1-\phi)/4, (1-\phi)/4, \phi/4),$$

  and $0 \leq \phi \leq 1$. Thus

  $$g(\mathrm{y}|\phi) = \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!} \cdot (\frac{1}{2}+\frac{\phi}{4})^{y_1}(\frac{1}{4}-\frac{\phi}{4})^{y_2}(\frac{1}{4}-\frac{\phi}{4})^{y_3}(\frac{\phi}{4})^{y_4}$$

  If we represent $y$ as incomplete data from a five-category multinomial with cellprobabilities

  $$(1/2, \phi/4, (1-\phi)/4, (1-\phi)/4, \phi/4).$$

Then the complete data is x $= (x_1, x_2, x_3, x_4, x_5)$ where $y_1 = x_1 + x_2$, $y_2 = x_3$, $y_3 = x_4$, $y_4 = x_5$. The complete data specification is

$$f(\mathrm{x}|\phi) = \frac{(x_1 + x_2 + x_3 + x_4 + x_5)!}{x_1!x_2!x_3!x_4!x_5!} \cdot \left(\frac{1}{2}\right)^{x_1} \left(\frac{\phi}{4}\right)^{x_2+x_5} \left(\frac{1}{4} - \frac{\phi}{4}\right)^{x_3+x_4}.$$

- E-step:
  estimate the sufficient statistics of the complete data $x$, given $y$:

$$x_1^{(p)} = 125\frac{\phi^{(p)}/4}{1/2 + \phi^{(p)}/4}$$
$$x_2^{(p)} = 125\frac{1/2}{1/2 + \phi^{(p)}/4}.$$

  Recall that $x_1 + x_2 = y_1 = 125$.

- M step:
  Maximizes the complete-data likelihood $f(x|\phi)$ with the estimated complete data $(x_1^{(p)}, x_2^{(p)}, 18, 20, 34)$. Thus we have

$$\pi^{(p+1)} = (x_2^{(p)} + 34)/(x_2^{(p)} + 72).$$

Estimate of $\phi$ is obtained by cycling back and forth between E- and M- steps.
This can be explicity solve a quadratic equation for the MLE of $\phi$:

$$\pi = \frac{\phi + 34}{\pi + 34 + 18 + 20},$$

with solution $\phi^* = 0.6268215$.

## 2. EM in exponential family

- $f(x|\phi) = b(x)\exp(\phi t(x)^T)/a(\phi)$,
  $\phi$: natural parameters
  $t(x)$: complete-data sufficient statistics for $\phi$.
  E-step:
  Estimate the complete-data sufficient statistics $t$ by

$$t^{(p)} = E(t|y, \phi^{(p)})$$

  $\phi^{(p)}$ is the current value of $\phi$ after $p$ cycles of the algorithm.
  M-step:
  Determine $\phi^{(p+1)}$ as

$$E(t(x)|\phi) = t^{(p)} = E(t|y, \phi^{(p)}).$$

  (which is the score equations given data from a regular exponential family)

-

$$\begin{aligned} k(x|y,\phi) &= f(x|\phi)/g(y|\phi) \\ &= b(x)\exp(\phi t^T(x))/a(\phi|y), \end{aligned}$$

where

$$a(\phi|y) = \int_{\mathcal{X}(y)} b(x) \exp(\phi t^T(x) dx.$$

- Note that

$$\begin{aligned} L(\phi) &= \log g(y|\phi) \\ &= \log f(x|\phi) - \log k(x|y,\phi) \\ &= -\log a(\phi) + \log a(\phi|y). \end{aligned}$$

$$\begin{aligned} DL(\phi) &= -D \log a(\phi) + D \log a(\phi|y) \\ &= -E(t|\phi) + E(t|y,\phi). \end{aligned}$$

Thus the score equations have an attractive representation as the difference of an unconditional and a conditional expectation of the sufficient statistics.

## 3. EM in general form

$$Q(\phi'|\phi) = E(\log f(x|\phi')|y, \phi)$$

- E-step: Compute $Q(\phi|\phi^{(p)})$.
- M-step: Choose $\phi^{(p+1)}$ to be a value of $\phi$ which maximizes $Q(\phi|\phi^{(p)})$.

(The heuristic idea is we would like to choose $\phi^*$ to maximize $\log f(x|\phi)$. Since we don't know $\log f(x|\phi)$, we maximize instead its current expectation given the observed data $y$ and the current fit $\phi^{(p)}$.)

## 4. General Properties

Write

$$H(\phi^{'}|\phi) = E(\log k(x|y,\phi^{'})|y,\phi).$$

$$(\Rightarrow Q(\phi^{'}|\phi) = L(\phi^{'}) + H(\phi^{'}|\phi))$$

Lemma 1. $H(\phi^{'}|\phi) \leq H(\phi|\phi)$, with equality iff $k(x|y,\phi^{'}) = k(x|y,\phi)$ a.e.

Def. An iterative algorithm with mapping $M(\phi)$ is a generalized EM algorithm (GEM) if

$$Q(M(\phi)|\phi) \geq Q(\phi|\phi)$$

for all $\phi \in \Omega$.

For EM, $Q(M(\phi)|\phi) \geq Q(\phi^{'}|\phi)$ for all $(\phi^{'},\phi)$ in $\Omega\times\Omega$, i.e., $\phi^{'} = M(\phi)$ maximizes $Q(\phi^{'}|\phi)$.)

Theorem 1. For every GEM,

$$L(M(\phi)) \geq L(\phi) \quad \text{for all } \phi \in \Omega,$$

with equality holds iff

$$Q(M(\phi)|\phi) = Q(\phi|\phi)$$

and

$$k(x|y,M(\phi)) = k(x|y,\phi) \text{ a.e..}$$

Proof. Observe that

$$\begin{aligned} L(M(\phi)) &\geq L(\phi) \\ &= \{Q(M(\phi)|\phi) - Q(\phi|\phi)\} \\ &\quad +\{H(\phi|\phi) - H(M(\phi)|\phi)\}. \end{aligned}$$

Theorem 1 implies that $L(\phi)$ is non-decreasing on each iteration of a GEM algorithm, and is strictly increasing on any iteration such that

$$Q(\phi^{(p+1)}|\phi^{(p)}) > Q(\phi^{(p)}|\phi^{(p)}).$$

Corollary 1. Suppose for some $\phi^* \in \Omega$, $L(\phi^*) \geq L(\phi)$ for all $\phi \in \Omega$. Then for every GEM,
(a) $L(M(\phi^*)) = L(\phi^*)$,
(b) $Q(M(\phi^*)|\phi^*) = Q(\phi^*|\phi^*)$,
(c) $k(x|y, M(\phi^*)) = k(x|y, \phi^*)$ a.e.
Corollary 2. If for some $\phi^* \in \Omega$, $L(\phi^*) > L(\phi)$ for all $\phi \in \Omega$ such that $\phi \neq \phi^*$, then for every GEM,

$$M(\phi^*) = \phi^*.$$

It states that a MLE is a fixed point of a GEM algorithm.

Theorem 2. Suppose that $\phi^{(p)}$ for $p = 0, 1, \ldots$ is an instance of a GEM algorithm such that
(1) the sequence $L(\phi^{(p)})$ is bounded, and
(2) $Q(\phi^{(p+1)}|\phi^{(p)}) - Q(\phi^{(p)}|\phi^{(p)})$
$\geq \lambda(\phi^{(p+1)} - \phi^{(p)})(\phi^{(p+1)} - \phi^{(p)})^T$ for some scalar $\lambda > 0$ and all $p$.
Then $\phi^{(p)}$ converges to some $\phi^*$ in the closure of $\Omega$.

Proof. Observe that

$$\begin{aligned} 0 &\leq Q(\phi^{(p+j)}|\phi^{(p+j-1)}) - Q(\phi^{(p+j-1)}|\phi^{(p+j-1)}) \\ &\geq L(\phi^{(p+j)}) - L(\phi^{(p+j-1)}). \end{aligned}$$

This implies that for all $r \geq 1$,

$$\lambda(\phi^{(p+r)} - \phi^{(p)})(\phi^{(p+r)} - \phi^{(p)})^T < \epsilon.$$

Lemma 2. For all $\phi \in \Omega$,

$$E\left[\frac{\partial}{\partial\phi}\log k(x|y,\phi)|y,\phi\right] = D^{10}H(\phi|\phi) = 0$$

and

$$\begin{aligned} V\left[\frac{\partial}{\partial\phi}\log k(x|y,\phi)|y,\phi\right] &= D^{11}H(\phi|\phi) \\ &= -D^{20}H(\phi|\phi). \end{aligned}$$

Proof. Recall that

$$H(\phi|\phi) = E\left[\log k(x|y,\phi)|y,\phi\right].$$

Differentiating $\int_{\mathcal{X}(y)} k(x|y,\phi)dx = 1$ under the integral sign.

Theorem 3. Suppose that $\phi^{(p)}$ for $p = 0, 1, \ldots$ is an instance of a GEM algorithm such that

$$D^{10}Q(\phi^{(p+1)}|\phi^{(p)}) = 0.$$

Then for all $p$, there exists a $\phi_0^{(p+1)} \in [\phi^{(p)}, \phi^{(p+1)}]$ such that

$$\begin{aligned} &Q(\phi^{(p+1)}|\phi^{(p)}) - Q(\phi^{(p)}|\phi^{(p)}) \\ &= -(\phi^{(p+1)} - \phi^{(p)})D^{20}Q(\phi_0^{(p+1)}|\phi^{(p)})(\phi^{(p+1)} - \phi^{(p)})^T. \end{aligned}$$

Further, if $D^{20}Q(\phi_0^{(p+1)}|\phi^{(p)})$ is negative definite, with eigenvalues bounded away from 0, and $L(\phi^{(p)})$ is bounded, then the sequence $\phi^{(p)}$ converges to some $\phi^{(*)}$ in the closure of $\Omega$.

Proof. Observe that

$$\begin{aligned} &Q(\phi|\phi^{(p)}) \approx Q(\phi^{(p+1)}|\phi^{(p)}) \\ &(\phi - \phi^{(p+1)})D^{10}Q(\phi^{(p+1)}|\phi^{(p)}) \\ &+(\phi - \phi^{(p+1)})D^{20}Q(\phi_0^{(p+1)}|\phi^{(p)})(\phi - \phi^{(p+1)})^T. \end{aligned}$$

Let $\phi = \phi^{(p)}$ we obtain the result. If $D^{20}Q(\phi_0^{(p+1)}|\phi^{(p)})$ are negative definite, with eigenvalues bounded

away from 0, then by condition (2) of Theorem 2 the sequence $\phi^{(p)}$ converges to some $\phi^{(*)}$ in the closure of $\Omega$.

Theorem 4. Suppose that $\phi^{(p)}$, $p = 0, 1, \ldots$ is an instance of a GEM algorithm such that
(1) $\phi^{(p)}$ converges to $\phi^*$ in the closure of $\Omega$.
(2) $D^{10}Q(\phi^{(p+1)}|\phi^{(p)}) = 0$ and
(3) $D^{20}Q(\phi^{(p+1)}|\phi^{(p)})$ is negative definite with eigenvalues bounded away from 0.
Then

- $DL(\phi^*) = 0$,
- $D^{20}Q(\phi^*|\phi^*)$ is nonnegative definite,
- $DM(\phi^*) = D^{20}H(\phi^*|\phi^*)[D^{20}Q(\phi^*|\phi^*)]^{-1}$.

Proof. (1) Observe that

$$DL(\phi^{(p+1)}) = -D^{10}H(\phi^{(p+1)}|\phi^{(p)}) + D^{10}Q(\phi^{(p+1)}|\phi^{(p)}).$$

$D^{10}H(\phi^{(p+1)}|\phi^{(p)}) = 0$ by assumption (2), while $-D^{10}Q(\phi^{(p+1)}|\phi^{(p)}) = 0$ in the limit as $p \to \infty$.
(2) $D^{20}Q(\phi^*|\phi^*)$ is nonnegative definite since it is the limit of $D^{20}Q(\phi^{(p+1)}|\phi^{(p)})$ whose eigenvalues are bounded away from zero.

(3) Note that

$$\begin{aligned} D^{10}Q(\phi_2|\phi_1) &= D^{10}Q(\phi^*|\phi^*) \\ &+(\phi_2-\phi^*)D^{20}Q(\phi^*|\phi^*) \\ &+(\phi_1-\phi^*)D^{11}Q(\phi^*|\phi^*)+\cdots. \end{aligned}$$

substituting $\phi_1 = \phi^{(p+1)}$ and $\phi_2 = \phi^{(p+1)}$,

$$\begin{aligned} 0 \;&=\; (\phi^{(p+1)}-\phi^*)D^{20}Q(\phi^*|\phi^*) \\ &+\; (\phi^{(t)}-\phi^*)D^{11}Q(\phi^*|\phi^*)+\cdots \end{aligned}$$

Since $\phi^{(p+1)} = M(\phi^{(p)})$ and $\phi^* = M(\phi^*)$, we have in the limit that

$$0 = DM(\phi^*)D^{20}Q(\phi^*|\phi^*) + D^{11}Q(\phi^*|\phi^*).$$

Note that

$$\begin{aligned} D^{11}Q(\phi^*|\phi^*) &= M^{(1)}(\phi^*) \\ &= -D^{20}H(\phi^*|\phi^*). \end{aligned}$$

When the eigenvalues of $DM(\phi^*)$ are all less than 1, the largest such eigenvalue gives the rate of convergence of the algorithm. Thus the rate of convergence depends on $-D^2L(\phi^*)$, a measure of the observed information, and $-D^{20}H(\phi^*|\phi^*)$, a measure of the missing information. When the information loss due to incompleteness is small,

the algorithm converges rapidly.
EX. For a regular exponential family,

$$D^{20}Q(\phi|\phi^{(p)}) = -V(t|\phi)$$

so that if the eigenvalues of $V(t|\phi)$ are bounded above zero, $\phi^{(p)}$ converges.
Note that

$$D^{20}H(\phi^*|\phi^*) = -V(t|y,\phi^*)$$

thus

$$DM(\phi^*) = V(t|y,\phi^*)[V(t|\phi^*)]^{-1}.$$

## 5. Applications of EM algorithms to incomplete data analysis

(1) Incomplete Covariate Data in Linear Regression Models:

- complete-data score

$$S(\beta) = \sum_i X_i(Y_i - X_i^T\beta)$$

- observed data: $(Y_i, Z_i)$
- score for the observed data

$$E[S(\beta)|Y, Z]$$
$$= \sum_i E_{\beta_1}(X_i|Y_i, Z_i)Y_i - \beta \sum_i E_{\beta_1}(X_i X_i^T|Y_i, Z_i)$$

To implement the EM, we need to specify, in addition to $P_\beta(Y|X, Z)$, a model for $P_\gamma(X|Z)$;

$$P_{\beta,\gamma}(X|Y, Z) = \frac{P_\beta(Y|X, Z)P_\gamma(X|Z)}{\int P_\beta(Y|X, Z)P_\gamma(X|Z)dX}.$$

(2) Surrogate Outcome Data in Linear Models:

- observed data: $(\tilde{Y}_i, X_i)$
  $\tilde{Y}$: surrogate for $Y$

- score for the observed data

$$E[S(\beta)|\tilde{Y}, X]$$
$$= \sum_i X_i\{E(Y|\tilde{Y}, X) - X_i^T\beta\}$$

To implement the EM, we need to specify, in addition to $P_\beta(Y|X)$, we need to specify a model for $P_\gamma(\tilde{Y}|Y, X)$;

$$P_{\beta,\gamma}(Y|\tilde{Y}, X) = \frac{P_\gamma(\tilde{Y}|Y, X)P_\beta(Y|X)}{\int P_\beta(\tilde{Y}|Y, X)P_\beta(Y|X)dY}.$$