Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2018 doi:10.20944/preprints201802.0131.v1

Tutorial on EM Algorithm

Loc Nguyen
Sunflower Soft Company, Vietnam
Email: ngphloc@sunflowersoft.net

Abstract
Maximum likelihood estimation (MLE) is a popular method for parameter estimation in both applied probability and statistics but MLE cannot solve the problem of incomplete data or hidden data because it is impossible to maximize likelihood function from hidden data. Expectation maximum (EM) algorithm is a powerful mathematical tool for solving this problem if there is a relationship between hidden data and observed data. Such hinting relationship is specified by a mapping from hidden data to observed data or by a joint probability between hidden data and observed data. In other words, the relationship helps us know hidden data by surveying observed data. The essential ideology of EM is to maximize the expectation of likelihood function over observed data based on the hinting relationship instead of maximizing directly the likelihood function of hidden data. Pioneers in EM algorithm proved its convergence. As a result, EM algorithm produces parameter estimators as well as MLE does. This tutorial aims to provide explanations of EM algorithm in order to help researchers comprehend it.
Keywords: expectation maximum, EM, generalized expectation maximum, GEM, EM convergence.

1. Introduction
We begin a review of expectation maximization (EM) algorithm with some basic concepts. Before discussing main subjects, there are some conventions. For example, if there is no additional explanation, random variables are denoted as uppercase letters such as X, Y, and Z. By default, vectors are column vectors. The tutorial is mainly extracted from the preeminent article “Maximum Likelihood from Incomplete Data via the EM Algorithm” by Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin (Dempster, Laird, & Rubin, 1977). For convenience, let **DLR** be reference to such three authors.

Suppose there are two samples $\boldsymbol{X}$ and $\boldsymbol{Y}$, in which $\boldsymbol{X}$ is *hidden space* (missing space) whereas $\boldsymbol{Y}$ is *observed space*. We do not know $\boldsymbol{X}$ but there is a mapping from $\boldsymbol{X}$ to $\boldsymbol{Y}$ so that we can survey $\boldsymbol{X}$ by observing $\boldsymbol{Y}$. The mapping is many-one function $\varphi: \boldsymbol{X} \rightarrow \boldsymbol{Y}$ and we denote $\boldsymbol{X}(Y)$ as all $X \in \boldsymbol{X}$ such that $\varphi(X) = Y$. So we have $\boldsymbol{X}(Y) = \{X: \varphi(X) = Y\}$. Let $f(X)$ be probability density function of random variable $X \in \boldsymbol{X}$ and let $g(Y)$ be probability density function of random variable $Y \in \boldsymbol{Y}$. Note, Y is also called observation. Equation 1.1 specifies $g(Y)$ as integral of $f(X)$ over $\boldsymbol{X}(Y)$.

$$g(Y|\Theta) = \int_{\boldsymbol{X}(Y)} f(X|\Theta)\mathrm{d}X \quad (1.1)$$

Where Θ is probabilistic parameter represented as a column vector, $\Theta = (\theta_1, \theta_2,\dots, \theta_r)^T$ in which each θ_i is a particular parameter with note that the superscript “T” denotes transposition operation for vector and matrix. Transposition operation transforms a column vector (column matrix) into a row vector (row matrix) and vice versa. Note that, Θ can simply be a scalar parameter. For example, normal distribution has two particular parameters such as mean μ and variance σ^2 and so we have $\Theta = (\mu, \sigma^2)^T$. The conditional probability density function of Y given X, denoted $k(X \mid Y, \Theta)$, is specified by equation 1.2.

$$k(X|Y,\Theta) = \frac{f(X|\Theta)}{g(Y|\Theta)} \quad (1.2)$$

© 2018 by the author(s). Distributed under a Creative Commons CC BY license.

DLR (Dempster, Laird, & Rubin, 1977, p. 1) considered ***X*** as *complete data* and ***Y*** as *incomplete data* because the mapping φ: ***X*** → ***Y*** is many-one function. Note that X and Y can be vectors or matrices but we survey they are scalar variables without loss of generality. In general, we only know ***Y*** and $f(X \mid \Theta)$ in order to determine $g(Y, \Theta)$ and $k(X \mid Y, \Theta)$. Our purpose is to estimate Θ based on such ***Y*** and $f(X \mid \Theta)$. Pioneers in expectation maximization (EM) algorithm firstly assumed that $f(X \mid \Theta)$ belongs to so-called exponential family with note that many popular distributions such as normal, multinomial, and Poisson belong to exponential family. Although DLR (Dempster, Laird, & Rubin, 1977) proposed a generality of EM algorithm in which $f(X \mid \Theta)$ distributes arbitrarily, we should concern exponential family a little bit. Exponential family (Wikipedia, Exponential family, 2016) refers to a set of probabilistic distributions whose density functions have the same exponential form according to equation 1.3 (Dempster, Laird, & Rubin, 1977, p. 3):

$$f(X|\Theta) = b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta) \quad (1.3)$$

Where $b(X)$ is a function of X, which is called base measure and $\tau(X)$ is a vector function of X, which is sufficient statistic. Let Ω be the convex set such that Θ ∈ Ω. If Θ is restricted only to Ω then, $f(X \mid \Theta)$ specifies a *regular exponential family*. If Θ lies in a curved sub-manifold of Ω then, $f(X \mid \Theta)$ specifies a *curved exponential family*. The $a(\Theta)$ is *partition function* for variable X, which is used to normalize PDF.

$$a(\Theta) = \int_X b(X)\exp\left(\Theta^T\tau(X)\right)\mathrm{d}X$$

The first-order derivative of $\log(a(\Theta))$ is expectation of $\tau(X)$.

$$\log'\left(a(\Theta)\right) = \frac{a'(\Theta)}{a(\Theta)} = \frac{\mathrm{d}\log\left(a(\Theta)\right)}{\mathrm{d}\Theta} = \frac{\mathrm{d}a(\Theta)/\mathrm{d}\Theta}{a(\Theta)} = \frac{1}{a(\Theta)}\frac{\mathrm{d}\left(\int_X b(X)\exp\left(\Theta^T\tau(X)\right)\mathrm{d}X\right)}{\mathrm{d}\Theta}$$
$$= \frac{1}{a(\Theta)}\int_X \frac{\mathrm{d}\left(b(X)\exp\left(\Theta^T\tau(X)\right)\right)}{\mathrm{d}\Theta}\mathrm{d}X = \int_X \tau(X)b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta)\,\mathrm{d}X$$
$$= E\left(\tau(X)|\Theta\right)$$

The second-order derivative of $\log(a(\Theta))$ is (Jebara, 2015):

$$\log''\left(a(\Theta)\right) = \frac{\mathrm{d}}{\mathrm{d}\Theta}\left(\frac{a'(\Theta)}{a(\Theta)}\right) = \frac{a''(\Theta)}{a(\Theta)} - \frac{a'(\Theta)}{a(\Theta)}\frac{\left(a'(\Theta)\right)^T}{a(\Theta)}$$
$$= \frac{a''(\Theta)}{a(\Theta)} - \left(E(\tau(X)|\Theta)\right)\left(E(\tau(X)|\Theta)\right)^T$$

Where,

$$\frac{a''(\Theta)}{a(\Theta)} = \frac{1}{a(\Theta)}\int_X \frac{\mathrm{d}^2\left(b(X)\exp\left(\Theta^T\tau(X)\right)\right)}{\mathrm{d}\Theta}\mathrm{d}X$$
$$= \int_X \left(\tau(X)\right)\left(\tau(X)\right)^T b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta)\,\mathrm{d}X = E\left(\left(\tau(X)\right)\left(\tau(X)\right)^T\middle|\Theta\right)$$

Hence (Hardle & Simar, 2013, pp. 125-126),

$$\log''\left(a(\Theta)\right) = E\left(\left(\tau(X)\right)\left(\tau(X)\right)^T\middle|\Theta\right) - \left(E(\tau(X)|\Theta)\right)\left(E(\tau(X)|\Theta)\right)^T = V(\tau(X)|\Theta)$$
$$= \int_X \left(\tau(X) - E(\tau(X)|\Theta)\right)\left(\tau(X) - E(\tau(X)|\Theta)\right)^T f(X|\Theta)\mathrm{d}X$$

Where $V(\tau(X) \mid \Theta)$ is central covariance matrix of $\tau(X)$. Please read the book "Matrix Analysis and Calculus" by Nguyen (Nguyen, 2015) for comprehending derivative of vector and matrix. Let $a(\Theta \mid Y)$ be a so-called *observed partition function* for observation Y.

$$a(\Theta|Y) = \int_{\boldsymbol{X}(Y)} b(X)\exp\left(\Theta^T\tau(X)\right)\mathrm{d}X$$

Similarly, we obtain that the first-order derivative of log(a($\boldsymbol{\theta}$ | Y)) is expectation of τ(X) based on Y.

$$\log'\left(a(\Theta|Y)\right) = \frac{1}{a(\Theta)}\frac{\mathrm{d}\left(\int_{\boldsymbol{X}(Y)} b(X)\exp\left(\Theta^T\tau(X)\right)\mathrm{d}X\right)}{\mathrm{d}\Theta} = E(\tau(X)|Y,\Theta)$$

If f(X | Θ) follows exponential family, the conditional density k(X | Y, Θ) is determined as follows:

$$k(X|Y,\Theta) = \frac{f(X|\Theta)}{g(Y|\Theta)}$$

If f(X | Θ) follows exponential family then, k(X | Y, Θ) also follows exponential family. In fact, we have:

$$k(X|Y,\Theta) = \frac{f(X|\Theta)}{g(Y|\Theta)} = \frac{b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta)}{\int_{\boldsymbol{X}(Y)} b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta)\,\mathrm{d}X} = \frac{b(X)\exp\left(\Theta^T\tau(X)\right)}{\int_{\boldsymbol{X}(Y)} b(X)\exp\left(\Theta^T\tau(X)\right)\mathrm{d}X}$$
$$= b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta|Y)$$

Note that k(X | Y, Θ) is determined on $X \in \boldsymbol{X}(Y)$. Of course, we have:

$$\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\mathrm{d}X = \int_{\boldsymbol{X}(Y)} \frac{b(X)\exp\left(\Theta^T\tau(X)\right)}{a(\Theta|Y)}\mathrm{d}X = \frac{\int_{\boldsymbol{X}(Y)} b(X)\exp\left(\Theta^T\tau(X)\right)\mathrm{d}X}{a(\Theta|Y)} = \frac{a(\Theta|Y)}{a(\Theta|Y)} = 1$$

The first-order derivative of log(a(Θ | Y)) is:

$$\log'\left(a(\Theta|Y)\right) = E(\tau(X)|Y,\Theta) = \int_X \tau(X)k(X|Y,\Theta)\mathrm{d}X$$

The second-order derivative of log(a(Θ) | Y) is:

$$\log''\left(a(\Theta|Y)\right) = V(\tau(X)|Y,\Theta)$$
$$= \int_X \left(\tau(X) - E(\tau(X)|Y,\Theta)\right)\left(\tau(X) - E(\tau(X)|Y,\Theta)\right)^T k(X|Y,\Theta)\mathrm{d}X$$

Where V(τ(X) | Y, Θ) is central covariance matrix of τ(X) given observed Y. Table 1.1 is summary of f(X | Θ), g(Y | Θ), k(X | Y, Θ), a(Θ), log'(a(Θ)), a(Θ | Y), and log'(a(Θ | Y)) with exponential family.

$$f(X|\Theta) = b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta)$$
$$g(Y|\Theta) = \int_{\boldsymbol{X}(Y)} b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta)\,\mathrm{d}X$$
$$k(X|Y,\Theta) = b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta|Y)$$
$$a(\Theta) = \int_X b(X)\exp\left(\Theta^T\tau(X)\right)\mathrm{d}X$$
$$\log'\left(a(\Theta)\right) = E(\tau(X)|\Theta)$$
$$\log''\left(a(\Theta)\right) = V(\tau(X)|\Theta)$$
$$a(\Theta|Y) = \int_{\boldsymbol{X}(Y)} b(X)\exp\left(\Theta^T\tau(X)\right)\mathrm{d}X$$
$$\log'\left(a(\Theta|Y)\right) = E(\tau(X)|Y,\Theta)$$
$$\log''\left(a(\Theta|Y)\right) = V(\tau(X)|Y,\Theta)$$
$$\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\mathrm{d}X = 1$$

Table 1.1. Summary of $f(X \mid \Theta)$, $g(Y \mid \Theta)$, $k(X \mid Y, \Theta)$, $a(\Theta)$, $\log'(a(\Theta))$, $a(\Theta \mid Y)$, and $\log'(a(\Theta \mid Y))$ with exponential family.

Simply, EM algorithm is iterative process including many iterations, in which each iteration has expectation step (E-step) and maximization step (M-step). E-step aims to estimate sufficient statistic given current parameter and observed data Y whereas M-step aims to re-estimate the parameter based on such sufficient statistic by maximizing likelihood function of X. EM algorithm is described in the next section in detail. As an introduction, DLR gave an example for illustrating EM algorithm (Dempster, Laird, & Rubin, 1977, pp. 2-3). Rao (Rao, 1955) presents observed data (incomplete data) Y of 197 animals following multinomial distribution with four categories, such as $Y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$. The probability density function of Y is:

$$g(Y|\theta) = \frac{\left(\sum_{i=1}^4 y_i\right)!}{\prod_{i=1}^4 y_i!} * \left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1} * \left(\frac{1}{4}-\frac{\theta}{4}\right)^{y_2} * \left(\frac{1}{4}-\frac{\theta}{4}\right)^{y_3} * \left(\frac{\theta}{4}\right)^{y_4}$$

Note, probabilities p_{y1}, p_{y2}, p_{y3}, and p_{y4} in $g(Y \mid \theta)$ are $1/2 + \theta/4$, $1/4 – \theta/4$, $1/4 – \theta/4$, and $\theta/4$, respectively as parameters. The expectation of any sufficient statistic y_i with regard to $g(Y \mid \theta)$ is:

$$E(y_i|Y,\theta) = y_i p_{y_i}$$

Observed data (incomplete data) Y is associated with hidden data X following multinomial distribution with five categories, such as $X = \{x_1, x_2, x_3, x_4, x_5\}$ where $y_1 = x_1 + x_2$, $y_2 = x_3$, $y_3 = x_4$, $y_4 = x_5$. The probability density function of X is:

$$f(X|\theta) = \frac{\left(\sum_{i=1}^5 x_i\right)!}{\prod_{i=1}^5 (x_i!)} * \left(\frac{1}{2}\right)^{x_1} * \left(\frac{\theta}{4}\right)^{x_2} * \left(\frac{1}{4}-\frac{\theta}{4}\right)^{x_3} * \left(\frac{1}{4}-\frac{\theta}{4}\right)^{x_4} * \left(\frac{\theta}{4}\right)^{x_5}$$

Note, probabilities p_{x1}, p_{x2}, p_{x3}, p_{x4}, and p_{x5} in $f(X \mid \theta)$ are $1/2$, $\theta/4$, $1/4 – \theta/4$, $1/4 – \theta/4$, and $\theta/4$, respectively as parameters. The expectation of any sufficient statistic x_i with regard to $f(X \mid \theta)$ is:

$$E(x_i|\theta) = x_i p_{x_i}$$

Due to $y_1 = x_1 + x_2$, $y_2 = x_3$, $y_3 = x_4$, $y_4 = x_5$, the mapping function φ between X and Y is $y_1 = \varphi(x_1, x_2) = x_1 + x_2$. Therefore $g(Y \mid \theta)$ is sum of $f(X \mid \theta)$ over x_1 and x_2 such that $x_1 + x_2 = y_1$ according to equation 1.1. In other words, $g(Y \mid \theta)$ is resulted from summing $f(X \mid \theta)$ over all (x_1, x_2) pairs such as (0, 125), (1, 124),…, (125, 0) because of $y_1 = 125$ from observed Y.

$$g(Y|\theta) = \sum_{x_1=0}^{125}\left(\sum_{x_2=125-x_1}^{0} f(X|\theta)\right)$$

Rao (Rao, 1955) applied EM algorithm into determining the optimal estimate θ^*. Note $y_2 = x_3$, $y_3 = x_4$, $y_4 = x_5$ are known and so only sufficient statistics x_1 and x_2 are not known. Given the t^{th} iteration, sufficient statistics x_1 and x_2 are estimated as $x_1^{(t)}$ and $x_2^{(t)}$ based on current parameter $\theta^{(t)}$ and $g(Y \mid \theta)$ in E-step below:

$$x_1^{(t)} + x_2^{(t)} = y_1^{(t)} = E\left(y_1 \middle| Y, \theta^{(t)}\right)$$

Due to $y_1 = 125$ from observed data and $p_{y1} = 1/2 + \theta/4$, which implies that:

$$x_1^{(t)} + x_2^{(t)} = E\left(y_1 \middle| Y, \theta^{(t)}\right) = y_1 p_{y_1} = 125\left(\frac{1}{2} + \frac{\theta^{(t)}}{4}\right)$$

We select:

$$x_1^{(t)} = 125\frac{1/2}{1/2 + \theta^{(t)}/4}$$
$$x_2^{(t)} = 125\frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4}$$

According to M-step, the next estimate $\theta^{(t+1)}$ is a maximizer of the log-likelihood function of X. This log-likelihood function is:

$$\log\big(f(X|\theta)\big) = \log\left(\frac{\left(\sum_{i=1}^{5} x_i\right)!}{\prod_{i=1}^{5}(x_i!)}\right) - (x_1 + 2x_2 + 2x_3 + 2x_4 + 2x_5)\log(2) + (x_2 + x_5)\log(\theta) + (x_3 + x_4)\log(1 - \theta)$$

The first-order derivative of $\log(f(X \mid \theta)$ is:

$$\frac{\mathrm{d}\log\big(f(X|\theta)\big)}{\mathrm{d}\theta} = \frac{x_2 + x_5}{\theta} - \frac{x_3 + x_4}{1 - \theta} = \frac{x_2 + x_5 - (x_2 + x_3 + x_4 + x_5)\theta}{\theta(1 - \theta)}$$

Because $y_2 = x_3 = 18$, $y_3 = x_4 = 20$, $y_4 = x_5 = 34$ and x_2 is approximated by $x_2^{(t)}$, we have:

$$\frac{\partial\log\big(f(X|\theta)\big)}{\partial\theta} = \frac{x_2^{(t)} + 34 - \left(x_2^{(t)} + 72\right)\theta}{\theta(1 - \theta)}$$

As a maximizer of $\log(f(X \mid \theta)$, the next estimate $\theta^{(t+1)}$ is solution of the following equation

$$\frac{\partial\log\big(f(X|\theta)\big)}{\partial\theta} = \frac{x_2^{(t)} + 34 - \left(x_2^{(t)} + 72\right)\theta}{\theta(1 - \theta)} = 0$$

So we have:

$$\theta^{(t+1)} = \frac{x_2^{(t)} + 34}{x_2^{(t)} + 72}$$

For example, given the initial $\theta^{(0)} = 0.5$, at the first iteration, we have:

$$x_2^{(1)} = 125\frac{\theta^{(0)}/4}{1/2 + \theta^{(0)}/4} = \frac{125 * 0.5/4}{0.5 + 0.5/4} = 25$$
$$\theta^{(1)} = \frac{x_2^{(1)} + 34}{x_2^{(1)} + 72} = \frac{25 + 34}{25 + 72} = 0.6082$$

After five iterations we gets the optimal estimate θ^*:

$$\theta^* = \theta^{(4)} = \theta^{(5)} = 0.6268$$

Table 1.2 (Dempster, Laird, & Rubin, 1977, p. 3) lists estimates of θ over four iterations (t =1, 2, 3, 4) with note that $\theta^{(0)}$ is initialized arbitrarily and $\theta^{(5)}$ is determined at the 4th iteration. The third column gives deviation $\theta^{(t)}$ and θ^* whereas the fourth column gives the ratio of successive deviations. Later on, we will know that such ratio implies convergence rate.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2018 doi:10.20944/preprints201802.0131.v1

t	$\theta^{(t)}$	$(\theta^{(t)} - \theta^*)$	$(\theta^{(t+1)} - \theta^*)$ / $(\theta^{(t)} - \theta^*)$
0	0.5	0.1268	0.1465
1	0.6082	0.0186	0.1346
2	0.6243	0.0025	0.1330
3	0.6265	0.0003	0.1328
4	0.6268	0	0.1328
5	0.6268	0	0.1328

Table 1.2. EM algorithm in simple case

2. EM algorithm

Expectation maximization (EM) algorithm has many iterations and each iteration has two steps in which expectation step (E-step) calculates sufficient statistic of hidden data based on observed data and current parameter whereas maximization step (M-step) re-estimates parameter. When DLR proposed EM algorithm (Dempster, Laird, & Rubin, 1977), they firstly concerned that the probability density function $f(X \mid \Theta)$ of hidden space belongs to exponential family. E-step and M-step at the t^{th} iteration are described in table 2.1 (Dempster, Laird, & Rubin, 1977, p. 4), in which the current estimate is $\Theta^{(t)}$.

E-step:

We calculate current value $\tau^{(t)}$ of the sufficient statistic $\tau(X)$ from observed Y and current parameter $\Theta^{(t)}$ as follows:

$$\tau^{(t)} = E\left(\tau(X) \middle| Y, \Theta^{(t)}\right)$$

M-step:

Basing on $\tau^{(t)}$, we determine the next parameter $\Theta^{(t+1)}$ as solution of following equation:

$$E\big(\tau(X)\big|\Theta\big) = \tau^{(t)}$$

Note, $\Theta^{(t+1)}$ will become current parameter at the next iteration ($(t+1)^{\text{th}}$ iteration).

Table 2.1. E-step and M-step of EM algorithm

EM algorithm stops if two successive estimates are equal, $\Theta^* = \Theta^{(t)} = \Theta^{(t+1)}$, at some t^{th} iteration. At that time we conclude that Θ^* is the optimal estimate of EM process. Please see table 1.1 to know how to calculate $E(\tau(X) \mid \Theta^{(t)})$ and $E(\tau(X) \mid Y, \Theta^{(t)})$.

It is necessary to explain E-step and M-step as well as convergence of EM algorithm. Essentially, the two steps aims to maximize log-likelihood function of Θ, denoted $L(\Theta)$, with respect to observation Y.

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}}\, L(\Theta)$$

Where,

$$L(\Theta) = \log\big(g(Y|\Theta)\big)$$

Note that log(.) denotes logarithm function. Therefore, EM algorithm is an extension of maximum likelihood estimation (MLE) method. In fact, let $l(\Theta)$ be log-likelihood function of Θ with respect to variable X.

$$l(\Theta) = \log\big(f(Y|\Theta)\big) = b(X)\exp\big(\Theta^T\tau(X)\big)/a(\Theta) = \log(X) + \Theta^T\tau(X) - \log\big(a(\Theta)\big)$$

By referring to table 1.1, the first-order derivative of $l(\Theta)$ is:

$$\frac{\mathrm{d}l(\Theta)}{\mathrm{d}\Theta} = \tau(X) - \log'\big(a(\Theta)\big) = \tau(X) - E\big(\tau(X)\big|\Theta\big)$$

Maximizing $l(\Theta)$ is to set the first-order derivative of $l(\Theta)$ to be zero. Therefore, the optimal estimate Θ_* is solution of the following equation which is specified in M-step.

$$E\big(\tau(X)\big|\Theta\big) = \tau(X)$$

The expression $E(\tau(X) \mid \Theta)$ is function of Θ but $\tau(X)$ is still dependent on X. Let $\tau^{(t)}$ be value of $\tau(X)$ at the t^{th} iteration of EM process, candidate for the best estimate of Θ is solution of equation 2.1 according to M-step.

$$E(\tau(X)|\Theta) = \tau^{(t)} \tag{2.1}$$

Thus, we will calculate $\tau^{(t)}$ by maximizing the log-likelihood function $L(\Theta)$ with respect to observation Y. Recall that maximizing $L(\Theta)$ is the ultimate purpose of EM algorithm.

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} L(\Theta)$$

Where,

$$L(\Theta) = \log\big(g(Y|\Theta)\big) = \log\left(\int_{X(Y)} f(X|\Theta)\mathrm{d}X\right) \tag{2.2}$$

Due to:

$$k(X|Y,\Theta) = \frac{f(X|\Theta)}{g(Y|\Theta)}$$

It implies:

$$L(\Theta) = \log\big(g(Y|\Theta)\big) = \log\big(f(X|\Theta)\big) - \log\big(k(X|Y,\Theta)\big)$$

Because $f(X \mid \Theta)$ belongs to exponential family, we have:

$$f(X|\Theta) = b(X)\exp\big(\Theta^T\tau(X)\big)/a(\Theta)$$
$$k(X|Y,\Theta) = b(X)\exp\big(\Theta^T\tau(X)\big)/a(\Theta|Y)$$

The log-likelihood function $L(\Theta)$ is reduced as follows:

$$L(\Theta) = -\log\big(a(\Theta)\big) + \log\big(a(\Theta|Y)\big)$$

By referring to table 1.1, the first-order derivative of $L(\Theta)$ is:

$$\frac{\mathrm{d}L(\Theta)}{\mathrm{d}\Theta} = -\log'\big(a(\Theta)\big) + \log'\big(a(\Theta|Y)\big) = -E\big(\tau(X)|\Theta\big) + E\big(\tau(X)|Y,\Theta\big)$$

Maximizing $L(\Theta)$ is to set the first-order derivative of $L(\Theta)$ to be zero as be zero as follows:

$$-E\big(\tau(X)|\Theta\big) + E\big(\tau(X)|Y,\Theta\big) = 0$$

It implies:

$$E\big(\tau(X)|\Theta\big) = E\big(\tau(X)|Y,\Theta\big)$$

Let $\Theta^{(t)}$ be the current estimate at some t^{th} iteration of EM process. Derived from the equality above, the value $\tau^{(t)}$ is calculated as seen in equation 2.3.

$$\tau^{(t)} = E\big(\tau(X)\big|Y,\Theta^{(t)}\big) \tag{2.3}$$

Equation 2.3 specifies the E-step of EM process. After t iterations we will obtain $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ such that $E(\tau(X) \mid Y, \Theta^{(t)}) = E(\tau(X) \mid Y, \Theta^*) = \tau^{(t)} = E(\tau(X) \mid \Theta^*) = E(\tau(X) \mid \Theta^{(t+1)})$ when $\Theta^{(t+1)}$ is solution of equation 2.1 (Dempster, Laird, & Rubin, 1977, p. 5). This means that Θ^* is the optimal estimate of EM process because Θ^* is solution of the equation:

$$E\big(\tau(X)|\Theta\big) = E\big(\tau(X)|Y,\Theta\big)$$

Thus, we conclude that Θ^* is the optimal estimate of EM process.

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} L(\Theta)$$

For further research, DLR gave a preeminent generality of EM algorithm (Dempster, Laird, & Rubin, 1977, pp. 6-11) in which $f(X \mid \Theta)$ specifies arbitrary distribution. In other words, there is no requirement of exponential family. They define the conditional expectation $Q(\Theta' \mid \Theta)$ according to equation 2.4 (Dempster, Laird, & Rubin, 1977, p. 6).

$$Q(\Theta'|\Theta) = E\Big(\log\big(f(X|\Theta')\big)\Big|Y,\Theta\Big) = \int_{X(Y)} k(X|Y,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X \tag{2.4}$$

The two steps of generalized EM (*GEM*) algorithm aims to maximize $Q(\Theta \mid \Theta^{(t)})$ at some t^{th} iteration as seen in table 2.2 (Dempster, Laird, & Rubin, 1977, p. 6).

E-step:

The expectation $Q(\Theta \mid \Theta^{(t)})$ is calculated based on current $\Theta^{(t)}$, according to equation 2.4. *M-step*: The next parameter $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$. Note that $\Theta^{(t+1)}$ will become current parameter at the next iteration ((*t*+1)$^{\text{th}}$ iteration).

Table 2.2. E-step and M-step of GEM algorithm

DLR proved that GEM algorithm converges at some t^{th} iteration. At that time, $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ is the optimal estimate of EM process. It is deduced from E-step and M-step that $Q(\Theta \mid \Theta^{(t)})$ is increased after every iteration. How to maximize $Q(\Theta \mid \Theta^{(t)})$ is optimization problem which is dependent on applications. For example, the popular method to solve optimization problem is Lagrangian duality (Jia, 2013). GEM algorithm still aims to maximize the log-likelihood function $L(\Theta)$ specified equation 2.2. The next section focuses on the convergence of GEM algorithm proved by DLR (Dempster, Laird, & Rubin, 1977, pp. 7-10) but firstly we should discuss some features of $Q(\Theta' \mid \Theta)$. In special case of exponential family, $Q(\Theta' \mid \Theta)$ is specified by equation 2.5.

$$Q(\Theta'|\Theta) = E\big(\log\big(b(X)\big)\big|Y,\Theta\big) + (\Theta')^T\tau_\Theta - \log\big(a(\Theta')\big) \tag{2.5}$$

Where,

$$E\big(\log\big(b(X)\big)\big|Y,\Theta\big) = \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(b(X)\big)\mathrm{d}X$$

$$\tau_\Theta = \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\tau(X)\mathrm{d}X$$

Following is a proof of equation 2.5.

$$Q(\Theta'|\Theta) = E\Big(\log\big(f(X|\Theta')\big)\Big|Y,\Theta\Big)$$

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\Big(b(X)\exp\big((\Theta')^T\tau(X)\big)/a(\Theta')\Big)\mathrm{d}X$$

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\Big(\log\big(b(X)\big) + (\Theta')^T\tau(X) - \log\big(a(\Theta')\big)\Big)\mathrm{d}X$$

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(b(X)\big)\mathrm{d}X + \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)(\Theta')^T\tau(X)\mathrm{d}X - \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(a(\Theta')\big)\mathrm{d}X$$

$$= E\big(\log\big(b(X)\big)\big|Y,\Theta\big) + (\Theta')^T\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\tau(X)\mathrm{d}X - \log\big(a(\Theta')\big)$$

$$= E\big(\log\big(b(X)\big)\big|Y,\Theta\big) + (\Theta')^T E\big(\tau(X)|Y,\Theta\big) - \log\big(a(\Theta')\big)$$

Because $k(X \mid Y, \Theta)$ belongs exponential family, the expectation $E(\tau(X) \mid Y, \Theta)$ is function of Θ, denoted τ_Θ. It implies:

$$Q(\Theta'|\Theta) = E\big(\log\big(b(X)\big)\big|Y,\Theta\big) + (\Theta')^T\tau_\Theta - \log\big(a(\Theta')\big)\blacksquare$$

If there is no mapping function φ: $\boldsymbol{X} \to \boldsymbol{Y}$, the equation 2.4 is modified with assumption that there is a joint probability of X and Y, denoted $P(X, Y \mid \Theta)$. Note that $P(X, Y \mid \Theta)$ can be discrete or continuous. The condition probability of X given Y is specified according to Bayes' rule as follows:

$$P(X|Y,\Theta) = \frac{P(X,Y|\Theta)}{\int_{X\in\boldsymbol{X}_0} P(X,Y|\Theta)\mathrm{d}X}$$

Note, $\boldsymbol{X}_0 \subseteq \boldsymbol{X}$ is domain of X. Given Y, we always have:

$$\int_{X\in\boldsymbol{X}_0} P(X|Y,\Theta)\mathrm{d}X = 1$$

Equation 2.6 specifies the conditional expectation $Q(\Theta' \mid \Theta)$ without mapping function.

$$Q(\Theta'|\Theta) = \int_{X \in \boldsymbol{X}_0} P(X|Y,\Theta)\log\big(P(X,Y|\Theta')\big)\mathrm{d}X \tag{2.6}$$

Note, the requirement of joint probability is stricter than requirement of mapping function φ and so, equation 2.4 is the most general definition of $Q(\Theta' \mid \Theta)$.

3. Convergence of EM algorithm

Recall that DLR proposed GEM algorithm which aims to maximize the log-likelihood function $L(\Theta)$ by maximizing $Q(\Theta' \mid \Theta)$ over many iterations. This section focuses on mathematical explanation of the convergence of GEM algorithm given by DLR (Dempster, Laird, & Rubin, 1977, pp. 6-9). Recall that we have:

$$L(\Theta) = \log\big(g(Y|\Theta)\big) = \log\left(\int_{\boldsymbol{X}(Y)} f(X|\Theta)\mathrm{d}X\right)$$

$$Q(\Theta'|\Theta) = E\Big(\log\big(f(X|\Theta')\big)\Big|Y,\Theta\Big) = \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X$$

Let $H(\Theta' \mid \Theta)$ be another conditional expectation which has strong relationship with $Q(\Theta' \mid \Theta)$ (Dempster, Laird, & Rubin, 1977, p. 6).

$$H(\Theta'|\Theta) = E\Big(\log\big(k(X|Y,\Theta')\big)\Big|Y,\Theta\Big) = \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta')\big)\mathrm{d}X \tag{3.1}$$

From equation 2.4 and equation 3.1, we have:

$$Q(\Theta'|\Theta) = L(\Theta') + H(\Theta'|\Theta) \tag{3.2}$$

Following is a proof of equation 3.2.

$$Q(\Theta'|\Theta) = \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X = \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(g(Y|\Theta')k(X|Y,\Theta')\big)\mathrm{d}X$$

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(g(Y|\Theta')\big)\mathrm{d}X + \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta')\big)\mathrm{d}X$$

$$= \log\big(g(Y|\Theta')\big)\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\mathrm{d}X + H(\Theta'|\Theta) = \log\big(g(Y|\Theta')\big) + H(\Theta'|\Theta)$$

$$= L(\Theta') + H(\Theta'|\Theta)\blacksquare$$

Lemma 1 (Dempster, Laird, & Rubin, 1977, p. 6). For any pair (Θ', Θ) in $\Omega \times \Omega$,

$$H(\Theta'|\Theta) \le H(\Theta|\Theta) \tag{3.3}$$

The equality occurs if and only if $k(X \mid Y, \Theta') = k(X \mid Y, \Theta)$ almost everywhere ■

Following is a proof of lemma 1 as well as equation 3.3. The log-likelihood function $L(\Theta')$ is re-written as follows:

$$L(\Theta') = \log\left(\int_{\boldsymbol{X}(Y)} f(X|\Theta')\mathrm{d}X\right) = \log\left(\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\frac{f(X|\Theta')}{k(X|Y,\Theta)}\mathrm{d}X\right)$$

Due to

$$\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta')\mathrm{d}X = 1$$

By applying Jensen's inequality (Sean, 2009, pp. 3-4) with concavity of logarithm function, Sean (Sean, 2009, p. 6) proved that:

$$L(\Theta') \ge \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\left(\frac{f(X|\Theta')}{k(X|Y,\Theta)}\right)\mathrm{d}X$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2018 doi:10.20944/preprints201802.0131.v1

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\left(\log\big(f(X|\Theta')\big) - \log\big(k(X|Y,\Theta)\big)\right)\mathrm{d}X$$
$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta')g(Y|\Theta')\big)\mathrm{d}X - \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta)\big)\mathrm{d}X$$
$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\left(\log\big(k(X|Y,\Theta')\big) + \log\big(g(Y|\Theta')\big)\right)\mathrm{d}X - H(\Theta|\Theta)$$
$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\left(\log\big(k(X|Y,\Theta')\big)\right)\mathrm{d}X + \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\left(\log\big(g(Y|\Theta')\big)\right)\mathrm{d}X - H(\Theta|\Theta)$$
$$= H(\Theta'|\Theta) + \log\big(g(Y|\Theta')\big)\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\mathrm{d}X - H(\Theta|\Theta)$$
$$= H(\Theta'|\Theta) + L(\Theta') - H(\Theta|\Theta)$$

It implies:

$$H(\Theta'|\Theta) \leq H(\Theta|\Theta)\blacksquare$$

According to Jensen's inequality (Sean, 2009, pp. 3-4), the equality occurs if and only if $k(X \mid Y, \Theta')$ is linear or $f(X \mid \Theta')$ is constant. In other words, the equality occurs if and only if $k(X \mid Y, \Theta') = k(X \mid Y, \Theta)$ almost everywhere when $f(X \mid \Theta)$ is not constant.

Let $\Theta^{(1)} \rightarrow \Theta^{(2)} \rightarrow \ldots \rightarrow \Theta^{(t)} \rightarrow \Theta^{(t+1)} \rightarrow \ldots$ be the sequence of estimates of Θ resulted from iterations of EM algorithm. Let $\Theta \rightarrow M(\Theta)$ be the mapping such that each estimation $\Theta^{(t)} \rightarrow \Theta^{(t+1)}$ at any given iteration is defined by equation 3.4 (Dempster, Laird, & Rubin, 1977, p. 7).

$$\Theta^{(t+1)} = M\big(\Theta^{(t)}\big) \tag{3.4}$$

Definition 1 (Dempster, Laird, & Rubin, 1977, p. 7). An iterative algorithm with mapping $M(\Theta)$ is a GEM algorithm if

$$Q(M(\Theta)|\Theta) \geq Q(\Theta|\Theta)\blacksquare \tag{3.5}$$

Of course, specification of GEM shown in table 2.2 satisfies the definition 1 because $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with regard to variable Θ in M-step.

$$Q\big(M\big(\Theta^{(t)}\big)\big|\Theta^{(t)}\big) = Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) \geq Q\big(\Theta^{(t)}\big|\Theta^{(t)}\big), \forall t$$

Theorem 1 (Dempster, Laird, & Rubin, 1977, p. 7). For every GEM algorithm

$$L\big(M(\Theta)\big) \geq L(\Theta) \text{ for all } \Theta \in \Omega \tag{3.6}$$

Where equality occurs if and only if $Q(M(\Theta) \mid \Theta) = Q(\Theta \mid \Theta)$ and $k(X \mid Y, M(\Theta)) = k(X \mid Y, \Theta)$ almost everywhere ■

Following is the proof of theorem 1 (Dempster, Laird, & Rubin, 1977, p. 7)

$$L\big(M(\Theta)\big) - L(\Theta) = \big(Q(M(\Theta)|\Theta) - H(M(\Theta)|\Theta)\big) - \big(Q(\Theta|\Theta) - H(\Theta|\Theta)\big)$$
$$= \big(Q(M(\Theta)|\Theta) - Q(\Theta|\Theta)\big) + \big(H(\Theta|\Theta) - H(M(\Theta)|\Theta)\big) \geq 0\blacksquare$$

Because the equality of lemma 1 occurs if and only if $k(X \mid Y, \Theta') = k(X \mid Y, \Theta)$ almost everywhere and the equality of the definition 1 is $Q(M(\Theta) \mid \Theta) = Q(\Theta \mid \Theta)$, we deduce that the equality of theorem 1 occurs if and only if $Q(M(\Theta) \mid \Theta) = Q(\Theta \mid \Theta)$ and $k(X \mid Y, M(\Theta)) = k(X \mid Y, \Theta)$ almost everywhere. It is easy to draw corollary 1 and corollary 2 from definition 1 and theorem 1.

Corollary 1 (Dempster, Laird, & Rubin, 1977). Suppose for some $\Theta^* \in \Omega$, $L(\Theta^*) \geq L(\Theta)$ for all $\Theta \in \Omega$ then for every GEM algorithm:

(a) $L(M(\Theta^*)) = L(\Theta^*)$
(b) $Q(M(\Theta^*) \mid \Theta^*) = Q(\Theta^* \mid \Theta^*)$
(c) $k(X \mid Y, M(\Theta^*)) = k(X \mid Y, \Theta^*)$ ■

Corollary 2 (Dempster, Laird, & Rubin, 1977). If for some $\Theta^* \in \Omega$, $L(\Theta^*) > L(\Theta)$ for all $\Theta \in \Omega$ such that $\Theta \neq \Theta^*$, then for every GEM algorithm:

$$M(\Theta^*) = \Theta^* \blacksquare$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2018 doi:10.20944/preprints201802.0131.v1

Theorem 2 (Dempster, Laird, & Rubin, 1977, p. 7). Suppose that $\Theta^{(t)}$ for t = 1, 2, 3,… is a sequence of estimates resulted from GEM algorithm such that:

(1) The sequence $L(\Theta^{(t)})$ is bounded, and

(2) $Q(\Theta^{(t+1)} \mid \Theta^{(t)}) - Q(\Theta^{(t)} \mid \Theta^{(t)}) \geq \xi(\Theta^{(t+1)} - \Theta^{(t)})^T(\Theta^{(t+1)} - \Theta^{(t)})$ for some scalar $\xi > 0$ and all t.

Then the sequence $\Theta^{(t)}$ converges to some Θ^* in the closure of Ω ■

Proof. The sequences $L(\Theta^{(t)})$ is non-decreasing according to theorem 1 and is bounded according to the assumption 1 of theorem 2 and hence, the sequence $L(\Theta^{(t)})$ converges to some $L^* < +\infty$. According to Cauchy criterion (Dinh, Pham, Nguyen, & Ta, 2000, p. 34), for all $\varepsilon > 0$, there exists a $t(\varepsilon)$ such that, for all $t \geq t(\varepsilon)$ and all $v \geq 1$:

$$L\left(\Theta^{(t+v)}\right) - L\left(\Theta^{(t)}\right) = \sum_{i=1}^{v}\left(L\left(\Theta^{(t+i)}\right) - L\left(\Theta^{(t+i-1)}\right)\right) < \varepsilon$$

By applying equations 3.2 and 3.3, for all $i \geq 1$, we obtain:

$$\begin{aligned}&Q\left(\Theta^{(t+i)}\middle|\Theta^{(t+i-1)}\right) - Q\left(\Theta^{(t+i-1)}\middle|\Theta^{(t+i-1)}\right)\\&= L\left(\Theta^{(t+i)}\right) + H\left(\Theta^{(t+i)}\middle|\Theta^{(t+i-1)}\right) - Q\left(\Theta^{(t+i-1)}\middle|\Theta^{(t+i-1)}\right)\\&\leq L\left(\Theta^{(t+i)}\right) + H\left(\Theta^{(t+i-1)}\middle|\Theta^{(t+i-1)}\right) - Q\left(\Theta^{(t+i-1)}\middle|\Theta^{(t+i-1)}\right)\\&= L\left(\Theta^{(t+i)}\right) - L\left(\Theta^{(t+i-1)}\right)\end{aligned}$$

It implies

$$\begin{aligned}\sum_{i=1}^{v}\left(Q\left(\Theta^{(t+i)}\middle|\Theta^{(t+i-1)}\right) - Q\left(\Theta^{(t+i-1)}\middle|\Theta^{(t+i-1)}\right)\right) &< \sum_{i=1}^{v}\left(L\left(\Theta^{(t+i)}\right) - L\left(\Theta^{(t+i-1)}\right)\right)\\&= L\left(\Theta^{(t+v)}\right) - L\left(\Theta^{(t)}\right) < \varepsilon\end{aligned}$$

By applying v times the assumption 2 of theorem 2, we obtain:

$$\begin{aligned}\varepsilon &> \sum_{i=1}^{v}\left(Q\left(\Theta^{(t+i)}\middle|\Theta^{(t+i-1)}\right) - Q\left(\Theta^{(t+i-1)}\middle|\Theta^{(t+i-1)}\right)\right)\\&\geq \xi\sum_{i=1}^{v}\left(\Theta^{(t+i)} - \Theta^{(t+i-1)}\right)^T\left(\Theta^{(t+i)} - \Theta^{(t+i-1)}\right)\end{aligned}$$

It means that

$$\sum_{i=1}^{v}\left|\Theta^{(t+i)} - \Theta^{(t+i-1)}\right|^2 < \varepsilon/\xi$$

Where,

$$\left|\Theta^{(t+i)} - \Theta^{(t+i-1)}\right|^2 = \left(\Theta^{(t+i)} - \Theta^{(t+i-1)}\right)^T\left(\Theta^{(t+i)} - \Theta^{(t+i-1)}\right)$$

Notation |.| denotes length of vector and so $|\Theta^{(t+i)} - \Theta^{(t+i-1)}|$ is distance between $\Theta^{(t+i)}$ and $\Theta^{(t+i-1)}$. Applying triangular inequality, for any $\varepsilon > 0$, for all $t \geq t(\varepsilon)$ and all $v \geq 1$, we have:

$$\left|\Theta^{(t+v)} - \Theta^{(t)}\right|^2 < \varepsilon/\xi$$

According to Cauchy criterion, the sequence $\Theta^{(t)}$ converges to some Θ^* in the closure of Ω.

Theorem 1 indicates that $L(\Theta)$ is non-decreasing on every iteration of GEM algorithm and is strictly increasing on any iteration such that $Q(\Theta^{(t+1)} \mid \Theta^{(t)}) > Q(\Theta^{(t)} \mid \Theta^{(t)})$. The corollaries 1 and 2 indicate that the optimal estimate is a fixed point of GEM algorithm. Theorem 2 points out convergence condition of GEM algorithm. However, there is still no assertion of convergence yet. The proof of convergence of GEM needs support of mathematical differentiation. There are two assumptions in this research:

1. There is a sufficient enough number of derivatives of $Q(\Theta' \mid \Theta)$, $L(\Theta)$, $H(\Theta' \mid \Theta)$, and $M(\Theta)$. In other words, $Q(\Theta' \mid \Theta)$, $L(\Theta)$, $H(\Theta' \mid \Theta)$, and $M(\Theta)$ are analytic functions (smooth functions).

2. Residuals of $Q(\Theta' \mid \Theta)$ and $H(\Theta' \mid \Theta)$ in second-order Taylor series are very small. Such residuals are negligible.

As a convention for derivatives of bivariate function, let D^{ij} denote as the derivative (differential) by taking i^{th}-order partial derivative with regard to first variable and then, taking j^{th}-order partial derivative with regard to second variable. If $i = 0$ ($j = 0$) then, there is no partial derivative with regard to first variable (second variable). For example, following is an example of how to calculate the derivative $D^{11}Q(\Theta^{(t)} \mid \Theta^{(t+1)})$.

- Firstly, we determine $D^{11}Q(\Theta'|\Theta) = \dfrac{\partial Q(\Theta'|\Theta)}{\partial\Theta'\partial\Theta}$
- Secondly, we substitute $\Theta^{(t)}$ and $\Theta^{(t+1)}$) for such $D^{11}Q(\Theta' \mid \Theta)$ to obtain $D^{11}Q(\Theta^{(t)} \mid \Theta^{(t+1)})$.

Equation 3.1 shows some derivatives (differentials) of $Q(\Theta' \mid \Theta)$, $H(\Theta' \mid \Theta)$, $L(\Theta)$, and $M(\Theta)$.

$$D^{10}Q(\Theta'|\Theta) = \frac{\partial Q(\Theta'|\Theta)}{\partial\Theta'}$$
$$D^{11}Q(\Theta'|\Theta) = \frac{\partial Q(\Theta'|\Theta)}{\partial\Theta'\partial\Theta}$$
$$D^{20}Q(\Theta'|\Theta) = \frac{\partial^2 Q(\Theta'|\Theta)}{\partial\Theta'\partial\Theta'}$$
$$D^{10}H(\Theta'|\Theta) = \frac{\partial H(\Theta'|\Theta)}{\partial\Theta'}$$
$$D^{11}H(\Theta'|\Theta) = \frac{\partial H(\Theta'|\Theta)}{\partial\Theta'\partial\Theta}$$
$$D^{20}H(\Theta'|\Theta) = \frac{\partial^2 H(\Theta'|\Theta)}{\partial\Theta'\partial\Theta'}$$
$$DL(\Theta) = \frac{\mathrm{d}L(\Theta)}{\mathrm{d}\Theta}$$
$$D^2L(\Theta) = \frac{\mathrm{d}^2L(\Theta)}{\mathrm{d}\Theta^2}$$
$$DM(\Theta) = \frac{\mathrm{d}M(\Theta)}{\mathrm{d}\Theta}$$

Table 3.1. Some derivatives of $Q(\Theta' \mid \Theta)$, $H(\Theta' \mid \Theta)$, $L(\Theta)$, and $M(\Theta)$

When Θ' and Θ are vectors, $D^{10}(\ldots)$ is gradient vector and $D^{20}(\ldots)$ is Hessian matrix. As a convention, let $\mathbf{0} = (0, 0,\ldots, 0)^T$ be zero vector.

Lemma 2 (Dempster, Laird, & Rubin, 1977, p. 8). For all Θ in Ω,

$$E\left(\frac{\mathrm{d}\log\big(k(X|Y,\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right) = D^{10}H(\Theta|\Theta) = \mathbf{0}^T \tag{3.7}$$

$$V_N\left(\frac{\mathrm{d}\log\big(k(X|Y,\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right) = D^{11}H(\Theta|\Theta) = -D^{20}H(\Theta|\Theta)$$
$$V_N\left(\frac{\mathrm{d}\log\big(k(X|Y,\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right) = E\left(\left(\frac{\mathrm{d}\log\big(k(X|Y,\Theta)\big)}{\mathrm{d}\Theta}\right)^2\middle|Y,\Theta\right) \tag{3.8}$$
$$D^{20}H(\Theta|\Theta) = E\left(\frac{d^2\log\big(k(X|Y,\Theta)\big)}{\mathrm{d}(\Theta)^2}\middle|Y,\Theta\right)$$

$$E\left(\frac{\mathrm{d}\log\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right) = D^{10}Q(\Theta|\Theta) = DL(\Theta) \tag{3.9}$$

$$V_N\left(\frac{\mathrm{d}\log\left(f(X|\Theta)\right)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)=D^2L(\Theta)+\left(DL(\Theta)\right)^2-D^{20}Q(\Theta|\Theta)$$
$$V_N\left(\frac{\mathrm{d}\log\left(f(X|\Theta)\right)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)=E\left(\left(\frac{\mathrm{d}\log\left(f(X|\Theta)\right)}{\mathrm{d}\Theta}\right)^2\middle|Y,\Theta\right)\tag{3.10}$$
$$D^{20}Q(\Theta|\Theta)=E\left(\frac{d^2\log\left(f(X|\Theta)\right)}{\mathrm{d}(\Theta)^2}\middle|Y,\Theta\right)$$

Note, $V_N(.)$ denotes non-central covariance matrix, which is covariance matrix. Followings are proofs of equations 3.7, 3.8, 3.9, and 3.10. In fact, we have:

$$D^{10}H(\Theta'|\Theta)=\frac{\partial}{\partial\Theta'}E\left(\log\left(k(X|Y,\Theta')\right)\middle|Y,\Theta\right)=\frac{\partial}{\partial\Theta'}\left(\int_{\boldsymbol{X}(Y)}k(X|Y,\Theta)\log\left(k(X|Y,\Theta')\right)\mathrm{d}X\right)$$
$$=\int_{\boldsymbol{X}(Y)}k(X|Y,\Theta)\frac{\mathrm{d}\log\left(k(X|Y,\Theta')\right)}{\mathrm{d}\Theta'}\mathrm{d}X=E\left(\frac{\mathrm{d}\log\left(k(X|Y,\Theta')\right)}{\mathrm{d}\Theta'}\middle|Y,\Theta\right)=$$
$$=\int_{\boldsymbol{X}(Y)}\frac{k(X|Y,\Theta)}{k(X|Y,\Theta')}\frac{\mathrm{d}\left(k(X|Y,\Theta')\right)}{\mathrm{d}\Theta'}\mathrm{d}X$$

It implies:

$$D^{10}H(\Theta|\Theta)=\int_{\boldsymbol{X}(Y)}\frac{k(X|Y,\Theta)}{k(X|Y,\Theta)}\frac{\mathrm{d}\left(k(X|Y,\Theta)\right)}{\mathrm{d}\Theta}\mathrm{d}X=\frac{\mathrm{d}}{\mathrm{d}\Theta}\left(\int_{\boldsymbol{X}(Y)}k(X|Y,\Theta)\mathrm{d}X\right)=\frac{\mathrm{d}}{\mathrm{d}\Theta}(1)=\mathbf{0}^T$$

We also have:

$$D^{11}H(\Theta'|\Theta)=\frac{\partial D^{10}H(\Theta'|\Theta)}{\partial\Theta}=\int_{\boldsymbol{X}(Y)}\frac{1}{k(X|Y,\Theta')}\frac{\mathrm{d}k(X|Y,\Theta)}{d\Theta}\frac{\mathrm{d}k(X|Y,\Theta')}{\mathrm{d}\Theta'}\mathrm{d}X$$

It implies:

$$D^{11}H(\Theta|\Theta)=\int_{\boldsymbol{X}(Y)}\frac{1}{k(X|Y,\Theta)}\frac{\mathrm{d}k(X|Y,\Theta)}{d\Theta}\frac{\mathrm{d}k(X|Y,\Theta)}{\mathrm{d}\Theta}\mathrm{d}X$$
$$=\int_{\boldsymbol{X}(Y)}k(X|Y,\Theta)\left(\frac{1}{k(X|Y,\Theta)}\frac{\mathrm{d}k(X|Y,\Theta)}{d\Theta}\right)^2\mathrm{d}X=V_N\left(\frac{\mathrm{d}\log\left(k(X|Y,\Theta)\right)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)$$

We also have:

$$D^{20}H(\Theta'|\Theta)=\frac{\partial D^{10}H(\Theta'|\Theta)}{\partial\Theta'}=E\left(\frac{d^2\log\left(k(X|Y,\Theta')\right)}{\mathrm{d}(\Theta')^2}\middle|Y,\Theta\right)$$
$$=-\int_{\boldsymbol{X}(Y)}\frac{k(X|Y,\Theta)}{\left(k(X|Y,\Theta')\right)^2}\left(\frac{\mathrm{d}k(X|Y,\Theta')}{\mathrm{d}\Theta'}\right)^2\mathrm{d}X=-E\left(\left(\frac{\mathrm{d}\log\left(k(X|Y,\Theta)\right)}{\mathrm{d}\Theta}\right)^2\middle|Y,\Theta\right)$$

It implies:

$$D^{20}H(\Theta|\Theta)=-\int_{\boldsymbol{X}(Y)}k(X|Y,\Theta)\left(\frac{1}{k(X|Y,\Theta)}\frac{\mathrm{d}k(X|Y,\Theta)}{d\Theta}\right)^2\mathrm{d}X$$
$$=-V_N\left(\frac{\mathrm{d}\log\left(k(X|Y,\Theta)\right)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)$$

We have:

$$D^{10}Q(\Theta'|\Theta) = \frac{\partial}{\partial\Theta'}\left(\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X\right) = \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\frac{\mathrm{d}\log\big(f(X|\Theta')\big)}{\mathrm{d}\Theta'}\mathrm{d}X$$

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\frac{\mathrm{d}\log\big(f(X|\Theta')\big)}{\mathrm{d}\Theta'}\mathrm{d}X = E\left(\frac{\mathrm{d}\log\big(f(X|\Theta')\big)}{\mathrm{d}\Theta'}\middle|Y,\Theta\right)$$

$$= \int_{\boldsymbol{X}(Y)} \frac{k(X|Y,\Theta)}{f(X|\Theta')}\frac{\mathrm{d}f(X|\Theta')}{\mathrm{d}\Theta'}\mathrm{d}X$$

It implies:

$$D^{10}Q(\Theta|\Theta) = \int_{\boldsymbol{X}(Y)} \frac{k(X|Y,\Theta)}{f(X|\Theta)}\frac{\mathrm{d}f(X|\Theta)}{\mathrm{d}\Theta}\mathrm{d}X = \int_{\boldsymbol{X}(Y)} \frac{1}{g(Y|\Theta)}\frac{\mathrm{d}f(X|\Theta)}{\mathrm{d}\Theta}\mathrm{d}X$$

$$= \frac{1}{g(Y|\Theta)}\int_{\boldsymbol{X}(Y)} \frac{\mathrm{d}f(X|\Theta)}{\mathrm{d}\Theta}\mathrm{d}X = \frac{1}{g(Y|\Theta)}\frac{\mathrm{d}}{\mathrm{d}\Theta}\left(\int_{\boldsymbol{X}(Y)} f(X|\Theta)\mathrm{d}X\right)$$

$$= \frac{1}{g(Y|\Theta)}\frac{\mathrm{d}g(Y|\Theta)}{\mathrm{d}\Theta} = \frac{\mathrm{d}\log\big(g(Y|\Theta)\big)}{\mathrm{d}\Theta} = DL(\Theta)$$

We have:

$$D^{20}Q(\Theta'|\Theta) = \frac{\partial D^{10}Q(\Theta'|\Theta)}{\partial\Theta'} = \frac{\partial}{\partial\Theta'}\left(\int_{\boldsymbol{X}(Y)} \frac{k(X|Y,\Theta)}{f(X|\Theta')}\frac{\mathrm{d}f(X|\Theta')}{\mathrm{d}\Theta'}\mathrm{d}X\right)$$

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\frac{d}{\mathrm{d}\Theta'}\left(\frac{\mathrm{d}f(X|\Theta')/\mathrm{d}\Theta'}{f(X|\Theta')}\right)\mathrm{d}X = E\left(\frac{d^2\log\big(f(X|\Theta')\big)}{\mathrm{d}(\Theta')^2}\middle|Y,\Theta\right)$$

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\left(\left(\mathrm{d}^2f(X|\Theta')/\mathrm{d}(\Theta')^2\right)f(X|\Theta') - \left(\mathrm{d}f(X|\Theta')/\mathrm{d}\Theta'\right)^2\right)/\big(f(X|\Theta')\big)^2\mathrm{d}X$$

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\frac{\left(\mathrm{d}^2f(X|\Theta')/\mathrm{d}(\Theta')^2\right)}{f(X|\Theta')}\mathrm{d}X - \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\left(\frac{\mathrm{d}f(X|\Theta')/\mathrm{d}\Theta'}{f(X|\Theta')}\right)^2\mathrm{d}X$$

$$= \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\frac{\left(\mathrm{d}^2f(X|\Theta')/\mathrm{d}(\Theta')^2\right)}{f(X|\Theta')}\mathrm{d}X - V_N\left(\frac{\mathrm{d}\log\big(f(X|\Theta')\big)}{\mathrm{d}\Theta'}\middle|Y,\Theta\right)$$

It implies:

$$D^{20}Q(\Theta|\Theta) = \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\frac{\left(\mathrm{d}^2f(X|\Theta)/\mathrm{d}(\Theta)^2\right)}{f(X|\Theta)}\mathrm{d}X - V_N\left(\frac{\mathrm{d}\log\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)$$

$$= \frac{1}{g(Y|\Theta)}\int_{\boldsymbol{X}(Y)} \frac{\mathrm{d}^2f(X|\Theta)}{\mathrm{d}(\Theta)^2}\mathrm{d}X - V_N\left(\frac{\mathrm{d}\log\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)$$

$$= \frac{1}{g(Y|\Theta)}\frac{\mathrm{d}^2}{\mathrm{d}(\Theta)^2}\left(\int_{\boldsymbol{X}(Y)} \frac{f(X|\Theta)}{\mathrm{d}\Theta}\mathrm{d}X\right) - V_N\left(\frac{\mathrm{d}\log\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)$$

$$= \frac{1}{g(Y|\Theta)}\frac{\mathrm{d}^2g(Y|\Theta)}{\mathrm{d}(\Theta)^2} - V_N\left(\frac{\mathrm{d}\log\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)$$

Due to:

$$D^2L(\Theta) = \frac{\mathrm{d}^2\log\big(g(Y|\Theta)\big)}{\mathrm{d}(\Theta)^2} = \frac{1}{g(Y|\Theta)}\frac{\mathrm{d}^2g(Y|\Theta)}{\mathrm{d}(\Theta)^2} - \big(DL(\Theta)\big)^2$$

We have:

$$D^{20}Q(\Theta|\Theta) = D^2L(\Theta) + \big(DL(\Theta)\big)^2 - V_N\left(\frac{\mathrm{d}\log\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right) \blacksquare$$

Theorem 3 (Dempster, Laird, & Rubin, 1977, p. 8). Suppose $\Theta^{(t)}$ where $t = 1, 2, 3,\ldots$ is an instance of GEM algorithm such that

$$D^{10}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) = \mathbf{0}^T$$

Then for all t:

$$Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) - Q\big(\Theta^{(t)}\big|\Theta^{(t)}\big) = -\frac{1}{2}\big(\Theta^{(t+1)} - \Theta^{(t)}\big)^T D^{20}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big)\big(\Theta^{(t+1)} - \Theta^{(t)}\big)$$

Furthermore, if the Hessian matrix $D^{20}Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is negative definite, and the sequence $L(\Theta^{(t)})$ is bounded then, the sequence $\Theta^{(t)}$ converges to some Θ^* in the closure of Ω ■

Note, if Θ is a scalar parameter, $D^{20}Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ degrades as a scalar and the concept "negative definite" becomes "negative" simply. Following is a proof of theorem 3.

Proof. Second-order Taylor series expending $Q(\Theta \mid \Theta^{(t)})$ at $\Theta = \Theta^{(t+1)}$ with very small residual to obtain:

$$Q\big(\Theta\big|\Theta^{(t)}\big) = Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) + D^{10}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big)\big(\Theta - \Theta^{(t+1)}\big)$$
$$+\frac{1}{2}\big(\Theta - \Theta^{(t+1)}\big)^T D^{20}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big)\big(\Theta - \Theta^{(t+1)}\big)$$
$$= Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) + \frac{1}{2}\big(\Theta - \Theta^{(t+1)}\big)^T D^{20}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big)\big(\Theta - \Theta^{(t+1)}\big)$$
$$\Big(\text{due to } D^{10}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) = \mathbf{0}^T\Big)$$

Let $\Theta = \Theta^{(t)}$, we have:

$$Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) - Q\big(\Theta^{(t)}\big|\Theta^{(t)}\big) = -\frac{1}{2}\big(\Theta^{(t+1)} - \Theta^{(t)}\big)^T D^{20}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big)\big(\Theta^{(t+1)} - \Theta^{(t)}\big)\blacksquare$$

If $D^{20}Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is negative definite then,

$$Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) - Q\big(\Theta^{(t)}\big|\Theta^{(t)}\big) = -\frac{1}{2}\big(\Theta^{(t+1)} - \Theta^{(t)}\big)^T D^{20}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big)\big(\Theta^{(t+1)} - \Theta^{(t)}\big)$$
$$> 0$$

When,

$$\big(\Theta^{(t+1)} - \Theta^{(t)}\big)^T\big(\Theta^{(t+1)} - \Theta^{(t)}\big) \geq 0$$

So there exists some $\xi > 0$ such that

$$Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) - Q\big(\Theta^{(t)}\big|\Theta^{(t)}\big) \geq \xi\big(\Theta^{(t+1)} - \Theta^{(t)}\big)^T\big(\Theta^{(t+1)} - \Theta^{(t)}\big)$$

In other words, the assumption 2 of theorem 2 is satisfied and hence, the sequence $\Theta^{(t)}$ converges to some Θ^* in the closure of Ω if the sequence $L(\Theta^{(t)})$ is bounded ■

Theorem 4 (Dempster, Laird, & Rubin, 1977, p. 9). Suppose $\Theta^{(t)}$ where $t = 1, 2, 3,\ldots$ is an instance of GEM algorithm such that

(1) $\Theta^{(t)}$ converges to Θ^* in the closure of Ω.

(2) $D^{10}Q(\Theta^{(t+1)} \mid \Theta^{(t)}) = \mathbf{0}^T$, for all t.

(3) $D^{20}Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is negative definite for all t.

Then $DL(\Theta^*) = \mathbf{0}^T$, $D^{20}Q(\Theta^* \mid \Theta^*)$ is negative definite, $D^{20}H(\Theta^{(t)} \mid \Theta^{(t)})$ is negative semi-definite for all t, and

$$DM(\Theta^*) = D^{20}H(\Theta^*|\Theta^*)\big(D^{20}Q(\Theta^*|\Theta^*)\big)^{-1}\blacksquare \tag{3.11}$$

The notation "–1" denotes inverse of matrix. Note, $DM(\Theta^*)$ is differential of $M(\Theta)$ at $\Theta = \Theta^*$, which implies convergence of GEM algorithm. Followings are proofs of theorem 4.

From equation 3.2, we have:

$$DL\big(\Theta^{(t+1)}\big) = D^{10}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) - D^{10}H\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) = -D^{10}H\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big)$$
$$\Big(\text{Due to } D^{10}Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) = \mathbf{0}^T\Big)$$

When t approaches $+\infty$ such that $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$ then, $D^{10}H(\Theta^* \mid \Theta^*)$ is zero according to equation 3.7 and so we have:

$$DL(\Theta^*) = \mathbf{0}^T$$

Of course, $D^{20}Q(\Theta^* \mid \Theta^*)$ is negative definite because $D^{20}Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is negative definite, when t approaches $+\infty$ such that $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$.

By second-order Taylor series expanding for $H(\Theta \mid \Theta^{(t)})$ at $\Theta = \Theta^{(t)}$ with very small residual, we have:

$$H\left(\Theta\middle|\Theta^{(t)}\right) = H\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) + D^{10}H\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\left(\Theta - \Theta^{(t)}\right) + \frac{1}{2}\left(\Theta - \Theta^{(t)}\right)^T D^{20}H\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\left(\Theta - \Theta^{(t)}\right)$$

$$= H\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) + \frac{1}{2}\left(\Theta - \Theta^{(t)}\right)^T D^{20}H\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\left(\Theta - \Theta^{(t)}\right)$$

(Due to $D^{10}H(\Theta^{(t)} \mid \Theta^{(t)}) = 0$ according to equation 3.7)

Due to $H(\Theta \mid \Theta^{(t)}) \le H(\Theta^{(t)} \mid \Theta^{(t)})$ by equation 3.3, we have:

$$\frac{1}{2}\left(\Theta - \Theta^{(t)}\right)^T D^{20}H\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\left(\Theta - \Theta^{(t)}\right) = H\left(\Theta\middle|\Theta^{(t)}\right) - H\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) \le 0$$

Hence, $D^{20}H(\Theta^{(t)} \mid \Theta^{(t)})$ is negative semi-definite for all t. I include the result "$D^{20}H(\Theta^{(t)} \mid \Theta^{(t)})$ is negative semi-definite for all t" in theorem 4 because the origin version of theorem 4 by DLR does not has this result. Of course, $D^{20}H(\Theta^* \mid \Theta^*)$ is negative semi-definite because $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ is negative semi-definite, when t approaches $+\infty$ such that $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$.

By first-order Taylor series expanding for $D^{10}Q(\Theta_2 \mid \Theta_1)$ as a function of Θ_1 at $\Theta_1 = \Theta^*$ with very small residual, we have:

$$D^{10}Q(\Theta_2|\Theta_1) = D^{10}Q(\Theta_2|\Theta^*) + (\Theta_1 - \Theta^*)^T D^{11}Q(\Theta_2|\Theta^*)$$

$Q(\Theta_2 \mid \Theta^*)$ is better approximated by $Q(\Theta^* \mid \Theta^*)$ with definition 1 of GEM algorithm to obtain:

$$D^{10}Q(\Theta_2|\Theta_1) = D^{10}Q(\Theta^*|\Theta^*) + (\Theta_1 - \Theta^*)^T D^{11}Q(\Theta^*|\Theta^*) \quad (3.12)$$

By first-order Taylor series expanding for $D^{10}Q(\Theta_2 \mid \Theta_1)$ as a function of Θ_2 at $\Theta_2 = \Theta^*$ with very small residual, we have:

$$D^{10}Q(\Theta_2|\Theta_1) = D^{10}Q(\Theta^*|\Theta_1) + (\Theta_2 - \Theta^*)^T D^{20}Q(\Theta^*|\Theta_1)$$

$Q(\Theta^* \mid \Theta_1)$ is better approximated by $Q(\Theta^* \mid \Theta^*)$ with definition 1 of GEM algorithm to obtain:

$$D^{10}Q(\Theta_2|\Theta_1) = D^{10}Q(\Theta^*|\Theta^*) + (\Theta_2 - \Theta^*)^T D^{20}Q(\Theta^*|\Theta^*) \quad (3.13)$$

By summing equation 12 and equation 13, we have:

$$2D^{10}Q(\Theta_2|\Theta_1) = 2D^{10}Q(\Theta^*|\Theta^*) + (\Theta_1 - \Theta^*)^T D^{11}Q(\Theta^*|\Theta^*) + (\Theta_2 - \Theta^*)^T D^{20}Q(\Theta^*|\Theta^*)$$

From assumption 2 of theorem 4, we have $D^{10}Q(\Theta^* \mid \Theta^*) = \mathbf{0}^T$ when t approaches $+\infty$ such that $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$. Hence,

$$2D^{10}Q(\Theta_2|\Theta_1) = (\Theta_1 - \Theta^*)^T D^{11}Q(\Theta^*|\Theta^*) + (\Theta_2 - \Theta^*)^T D^{20}Q(\Theta^*|\Theta^*)$$

Substituting $\Theta_1 = \Theta^{(t)}$ and $\Theta_2 = \Theta^{(t+1)}$, we have:

$$2D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) = \left(\Theta^{(t)} - \Theta^*\right)^T D^{11}Q(\Theta^*|\Theta^*) + \left(\Theta^{(t+1)} - \Theta^*\right)^T D^{20}Q(\Theta^*|\Theta^*)$$

Due to $D^{10}Q(\Theta^{(t+1)} \mid \Theta^{(t)}) = \mathbf{0}^T$ then, we have:

$$\left(\Theta^{(t)} - \Theta^*\right)^T D^{11}Q(\Theta^*|\Theta^*) + \left(\Theta^{(t+1)} - \Theta^*\right)^T D^{20}Q(\Theta^*|\Theta^*) = \mathbf{0}^T$$

It implies:

$$\left(\Theta^{(t+1)} - \Theta^*\right)^T = \left(\Theta^{(t)} - \Theta^*\right)^T \left(-D^{11}Q(\Theta^*|\Theta^*)\left(D^{20}Q(\Theta^*|\Theta^*)\right)^{-1}\right)$$

In other words, we have:

$$\left(M\left(\Theta^{(t)}\right) - M(\Theta^*)\right)^T = \left(\Theta^{(t)} - \Theta^*\right)^T \left(-D^{11}Q(\Theta^*|\Theta^*)\left(D^{20}Q(\Theta^*|\Theta^*)\right)^{-1}\right)$$

When t approaches $+\infty$, we obtain $DM(\Theta^*)$ as differential of $M(\Theta)$ at Θ^*:

$$DM(\Theta^*) = -D^{11}Q(\Theta^*|\Theta^*)\left(D^{20}Q(\Theta^*|\Theta^*)\right)^{-1} \quad (3.14)$$

The derivative $D^{11}Q(\Theta' \mid \Theta)$ is expended as follows:

$$D^{11}Q(\Theta'|\Theta) = DL(\Theta') + D^{11}H(\Theta'|\Theta)$$

It implies:
$D^{11}Q(\Theta^*|\Theta^*) = DL(\Theta^*) + D^{11}H(\Theta^*|\Theta^*)$
$= 0 + D^{11}H(\Theta^*|\Theta^*)$
(Due to theorem 4)
$= -D^{20}H(\Theta^*|\Theta^*)$
(Due to equation 3.9)
Therefore, equation 3.14 becomes equation 3.11.

$$DM(\Theta^*) = D^{20}H(\Theta^*|\Theta^*)\left(D^{20}Q(\Theta^*|\Theta^*)\right)^{-1} \blacksquare$$

Finally, theorem 4 is proved. Because we assumed that $Q(\Theta' \mid \Theta)$ has second-order derivative, if $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ in M-step then, we always have $D^{10}Q(\Theta^{(t+1)} \mid \Theta^{(t)}) = 0$ as the assumption of theorem 3. By combination of theorems 3 and 4, we assert convergence of GEM, according to corollary 3.

Corollary 3. If an algorithm satisfies definition 1 such that $Q(M(\Theta) \mid \Theta) \geq Q(\Theta \mid \Theta^{(t)})$ then such algorithm is an GEM and converges to a stationary point Θ^* of the likelihood function $L(\Theta)$ such that $DL(\Theta^*) = \mathbf{0}^T$ ■

According to corollary 3, we do not assert whether Θ^* is a maximizer of $L(\Theta)$ yet. In a worst case, Θ^* may be a saddle point of $L(\Theta)$. Wu (Wu, 1983) answered exactly the question "Is Θ^* local maximizer, global maximizer, or saddle point?" in her/his article "On the Convergence Properties of the EM Algorithm".

Recall that $L(\Theta)$ is the log-likelihood function of observed Y according to equation 2.2.

$$L(\Theta) = \log\left(g(Y|\Theta)\right) = \log\left(\int_{\boldsymbol{X}(Y)} f(X|\Theta)\mathrm{d}X\right)$$

By default, in this study research, $D^{20}Q(\Theta \mid \Theta^{(t)})$ is assumed to be negative semi-definite for all $\Theta \in \Omega$.

Both $-D^{20}H(\Theta^* \mid \Theta^*)$ and $-D^{20}Q(\Theta^* \mid \Theta^*)$ are Fisher information matrices (Zivot, 2009, pp. 7-9) specified by equation 3.15.

$$\begin{aligned} I_H(\Theta^*) &= -D^{20}H(\Theta^*|\Theta^*) \\ I_Q(\Theta^*) &= -D^{20}Q(\Theta^*|\Theta^*) \end{aligned} \quad (3.15)$$

$I_H(\Theta^*)$ measures information of X about Θ^* with support of Y whereas $I_Q(\Theta^*)$ measures information of X about Θ^*. In other words, $I_H(\Theta^*)$ measures observed information whereas $I_Q(\Theta^*)$ measures hidden information. Let $V_H(\Theta^*)$ and $V_Q(\Theta^*)$ be covariance matrices of Θ^* with regard to $I_H(\Theta^*)$ and $I_Q(\Theta^*)$, respectively. They are inverses of $I_H(\Theta^*)$ and $I_Q(\Theta^*)$, according to equation 3.16.

$$\begin{aligned} V_H(\Theta^*) &= \left(I_H(\Theta^*)\right)^{-1} \\ V_Q(\Theta^*) &= \left(I_Q(\Theta^*)\right)^{-1} \end{aligned} \quad (3.16)$$

Equation 3.17 is a variant of equation 3.11 to calculate $DM(\Theta^*)$ based on information matrices:

$$DM(\Theta^*) = I_H(\Theta^*)\left(I_Q(\Theta^*)\right)^{-1} = \left(V_H(\Theta^*)\right)^{-1}V_Q(\Theta^*) \quad (3.17)$$

If $f(X \mid \Theta)$, $g(Y \mid \Theta)$, and $k(X \mid Y, \Theta)$ belong to exponential family, we have:

$$\frac{\mathrm{d}^2\log\left(f(Y|\Theta)\right)}{\mathrm{d}\Theta^2} = \frac{\mathrm{d}}{\mathrm{d}\Theta^2}\left(b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta)\right) = -\log''\left(a(\Theta)\right) = -V\left(\tau(X)|\Theta\right)$$

And

$$\frac{\mathrm{d}^2\log\left(k(X|Y,\Theta)\right)}{\mathrm{d}\Theta^2} = \frac{\mathrm{d}}{\mathrm{d}\Theta^2}\left(b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta|Y)\right) = -\log''\left(a(\Theta|Y)\right)$$
$$= -V\left(\tau(X)|Y,\Theta\right)$$

Please see table 1.1 to understand $V(\tau(X) \mid \Theta)$ and $V(\tau(X) \mid \Theta)$. With exponential family, we deduce that

$$D^{20}H(\Theta'|\Theta) = \int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\frac{\mathrm{d}^2\log\big(k(X|Y,\Theta')\big)}{\mathrm{d}(\Theta')^2}\mathrm{d}X = -\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\log''\big(a(\Theta')\big)\mathrm{d}X$$
$$= -\log''\big(a(\Theta')\big)\int_{\boldsymbol{X}(Y)} k(X|Y,\Theta)\mathrm{d}X = -\log''\big(a(\Theta')\big) = -V(\tau(X)|Y,\Theta')$$

Similarly, we have:

$$D^{20}Q(\Theta'|\Theta) = -V(\tau(X)|Y,\Theta')$$

Hence, equation 3.18 specifies $DM(\Theta^*)$ in case of exponential family.

$$DM(\Theta^*) = V(\tau(X)|Y,\Theta^*)V(\tau(X)|\Theta^*) \quad (3.18)$$

Suppose all partial derivatives of $H(\Theta' \mid \Theta)$ and $Q(\Theta' \mid \Theta)$ are continuous so that $D^{20}H(\Theta^* \mid \Theta^*)$ and $D^{20}Q(\Theta^* \mid \Theta^*)$ are symmetric matrices according to Schwarz's theorem (Wikipedia, Symmetry of second derivatives, 2018). Thus, $D^{20}H(\Theta^* \mid \Theta^*)$ and $D^{20}Q(\Theta^* \mid \Theta^*)$ are commutative:

$$D^{20}H(\Theta^* \mid \Theta^*)D^{20}Q(\Theta^* \mid \Theta^*) = D^{20}Q(\Theta^* \mid \Theta^*)H^{20}Q(\Theta^* \mid \Theta^*)$$

Suppose both $D^{20}H(\Theta^* \mid \Theta^*)$ and $D^{20}Q(\Theta^* \mid \Theta^*)$ are diagonalizable then, they are simultaneously diagonalizable (Wikipedia, Commuting matrices, 2017). Hence there is a (orthogonal) eigenvector matrix U such that (Wikipedia, Diagonalizable matrix, 2017) (StackExchange, 2013):

$$D^{20}H(\Theta^*|\Theta^*) = UH_e^*U^{-1}$$
$$D^{20}Q(\Theta^*|\Theta^*) = UQ_e^*U^{-1}$$

Where H_e^* and Q_e^* are eigenvalue matrices of $D^{20}H(\Theta^* \mid \Theta^*)$ and $D^{20}Q(\Theta^* \mid \Theta^*)$, respectively, according to equations 3.19 and 3.20. Of course, $h_1^*, h_2^*,\ldots, h_r^*$ are eigenvalues of $D^{20}H(\Theta^* \mid \Theta^*)$ whereas $q_1^*, q_2^*,\ldots, q_r^*$ are eigenvalues of $D^{20}Q(\Theta^* \mid \Theta^*)$.

$$H_e^* = \begin{pmatrix} h_1^* & 0 & \cdots & 0 \\ 0 & h_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_r^* \end{pmatrix} \quad (3.19)$$

$$Q_e^* = \begin{pmatrix} q_1^* & 0 & \cdots & 0 \\ 0 & q_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_r^* \end{pmatrix} \quad (3.20)$$

From equation 3.11, $DM(\Theta^*)$ is decomposed as seen in equation 3.21.

$$\begin{aligned} DM(\Theta^*) &= (UH_e^*U^{-1})(UQ_e^*U^{-1})^{-1} = UH_e^*U^{-1}U(Q_e^*)^{-1}\Lambda^{-1}U^{-1} \\ &= U(H_e^*(Q_e^*)^{-1})U^{-1} \end{aligned} \quad (3.21)$$

Let M_e^* be eigenvalue matrix of $DM(\Theta^*)$, specified by equation 22. As a convention M_e^* is called convergence matrix.

$$M_e^* = H_e^*(Q_e^*)^{-1} = \begin{pmatrix} m_1^* = \dfrac{h_1^*}{q_1^*} & 0 & \cdots & 0 \\ 0 & m_2^* = \dfrac{h_2^*}{q_2^*} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_r^* = \dfrac{h_r^*}{q_r^*} \end{pmatrix} \quad (3.22)$$

Of course, all $m_i^* = h_i^* / q_i^*$ are eigenvalues of $DM(\Theta^*)$. From theorem 4, $D^{20}Q(\Theta^* \mid \Theta^*)$ is negative definite and $D^{20}H(\Theta^* \mid \Theta^*)$ is negative semi-definite, which means that $q_i^* < 0$, $h_i^* \leq 0$, and $m_i^* \geq 0$, according equation 3.23.

$$q_i^* < 0, \forall i$$
$$h_i^* \leq 0, \forall i \tag{3.23}$$
$$m_i^* \geq 0, \forall i$$

In general, $D^{20}Q(\Theta^* \mid \Theta^*)$ and Q_e^* are negative definite, $D^{20}H(\Theta^* \mid \Theta^*)$ and H_e^* are negative semi-definite, and $DM(\Theta^*)$ and M_e^* are positive semi-definite.

Suppose $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)},\ldots, \theta_r^{(t)})$ at current t^{th} iteration and $\Theta^* = (\theta_1^*, \theta_2^*,\ldots, \theta_r^*)$, each m_i^* measures how much the next $\theta_i^{(t+1)}$ is near to θ_i^*. In other words, the smaller the m_i^* (s) are, the faster the GEM is and so the better the GEM is. This is why DLR (Dempster, Laird, & Rubin, 1977, p. 10) defined that the convergence rate m^* of GEM is the maximum one among all m_i^*, as seen in equation 3.24. The convergence rate m^* implies lowest speed.

$$m^* = \max_{m_i^*}\{m_1^*, m_2^*, \ldots, m_r^*\} \text{ where } m_1^* = \frac{h_1^*}{q_1^*} \tag{3.24}$$

From equations 3.2 and 3.11, we have (Dempster, Laird, & Rubin, 1977, p. 10):

$$D^2L(\Theta^*) = D^{20}Q(\Theta^*|\Theta^*) - D^{20}H(\Theta^*|\Theta^*) = D^{20}Q(\Theta^*|\Theta^*) - D^{20}Q(\Theta^*|\Theta^*)DM(\Theta^*)$$
$$= D^{20}Q(\Theta^*|\Theta^*)\big(I - DM(\Theta^*)\big)$$

Where I is identity matrix:

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

By the same way to draw convergence matrix M_e^* with note that $D^{20}H(\Theta^* \mid \Theta^*)$, $D^{20}Q(\Theta^* \mid \Theta^*)$, and $DM(\Theta^*)$ are symmetric matrices, we have:

$$L_e = Q_e(I - M_e) \tag{3.25}$$

Where L_e^* is eigenvalue matrix of $D^2L(\Theta^*)$. From equation 3.25, each eigenvalue l_i^* of L_e^* is proportional to each eigenvalues q_i^* of Q_e^* with ratio $1–m_i^*$ where m_i^* is an eigenvalue of M_e^*. Equation 3.26 specifies a so-called speed matrix S_e^*:

$$S_e^* = \begin{pmatrix} s_1^* = 1 - m_1^* & 0 & \cdots & 0 \\ 0 & s_2^* = 1 - m_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_r^* = 1 - m_r^* \end{pmatrix} \tag{3.26}$$

Equation 3.27 specifies L_e^* which is eigenvalue matrix of $D^2L(\Theta^*)$.

$$L_e^* = \begin{pmatrix} l_1^* = q_1^* s_1^* & 0 & \cdots & 0 \\ 0 & l_2^* = q_2^* s_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_r^* = q_r^* s_r^* \end{pmatrix} \tag{3.27}$$

From equation 3.26, suppose $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)},\ldots, \theta_r^{(t)})$ at current t^{th} iteration and $\Theta^* = (\theta_1^*, \theta_2^*,\ldots, \theta_r^*)$, each $s_i^* = 1–m_i^*$ is really the speed that the next $\theta_i^{(t+1)}$ moves to θ_i^*. From equations 3.24 and 3.26, equation 3.28 specifies the speed s^* of GEM algorithm.

$$s^* = 1 - m^*$$

Where,

$$m^* = \max_{m_i^*}\{m_1^*, m_2^*, \ldots, m_r^*\} \tag{3.28}$$

As a convention, if GEM algorithm fortunately stops at the first iteration such that $\Theta^{(1)} = \Theta^{(2)} = \Theta^*$ then, $m^* = 0$ and $s^* = 1$.

Without loss of generality, suppose parameter Θ degrades into scalar as $\Theta = \theta$ then, equation 3.11 is re-written as equation 3.29:

$$DM(\theta^*) = M_e^* = m^* = \lim_{t\to+\infty}\frac{M\big(\theta^{(t)}\big) - M(\theta^*)}{\theta^{(t)} - \theta^*} = \lim_{t\to+\infty}\frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*} = \tag{3.29}$$
$$= D^{20}H(\theta^*|\theta^*)\big(D^{20}Q(\theta^*|\theta^*)\big)^{-1}$$

A variant of equation 3.29 is equation 3.30 (McLachlan & Krishnan, 1997, p. 120).

$$DM(\theta^*) = M_e = m^* = \lim_{t \to +\infty} \frac{\theta^{(t+2)} - \theta^{(t+1)}}{\theta^{(t+1)} - \theta^{(t)}} \quad (3.30)$$

Proof. From equation 3.29, the next estimate $\theta^{(t+1)}$ approaches θ^* when $t \to +\infty$ and so we have:

$$DM(\theta^*) = M_e^* = m^* = \lim_{t \to +\infty} \frac{M\left(\theta^{(t)}\right) - M\left(\theta^{(t+1)}\right)}{\theta^{(t)} - \theta^{(t+1)}} = \lim_{t \to +\infty} \frac{\theta^{(t+1)} - \theta^{(t+2)}}{\theta^{(t)} - \theta^{(t+1)}}$$

$$= \lim_{t \to +\infty} \frac{\theta^{(t+2)} - \theta^{(t+1)}}{\theta^{(t+1)} - \theta^{(t)}}$$

Because the sequence $L(\theta^{(1)})$, $L(\theta^{(2)})$,…, $L(\theta^{(t)})$ is non-decreasing, the sequence $\theta^{(1)}$, $\theta^{(2)}$,…, $\theta^{(t)}$ is monotonous. This means:

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_t \leq \theta_{t+1} \leq \cdots \leq \theta^*$$

Or

$$\theta_1 \geq \theta_2 \geq \cdots \geq \theta_t \geq \theta_{t+1} \geq \cdots \geq \theta^*$$

It implies

$$0 \leq \frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*} \leq 1$$

So we have

$$0 \leq DM(\theta^*) = M_e^* = \lim_{t \to +\infty} \frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*} \leq 1$$

The fourth column of table 1.1 (Dempster, Laird, & Rubin, 1977, p. 3) gives sequences which approaches $M_e^* = DM(\theta^*)$ through many iterations by the following ratio to determine the limit in equation 3.29 with $\theta^* = 0.6268$.

$$\frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*}$$

In practice, if GEM is run step by step, θ^* is not known yet at some t^{th} iteration when GEM does not converge yet. Hence, equation 3.30 (McLachlan & Krishnan, 1997, p. 120) is used to make approximation of $M_e^* = DM(\theta^*)$ with unknown θ^* and $\theta^{(t)} \neq \theta^{(t+1)}$.

$$DM(\theta^*) \approx \frac{\theta^{(t+2)} - \theta^{(t+1)}}{\theta^{(t+1)} - \theta^{(t)}}$$

It is required only two successive iterations because both $\theta^{(t)}$ and $\theta^{(t+1)}$ are determined at t^{th} iteration whereas $\theta^{(t+2)}$ is determined at $(t+1)^{\text{th}}$ iteration. For example, in table 1.1, given $\theta^{(1)} = 0.6082$, $\theta^{(2)} = 0.6243$, and $\theta^{(3)} = 0.6265$, at $t = 1$, we have:

$$DM(\theta^*) \approx \frac{\theta^{(3)} - \theta^{(2)}}{\theta^{(2)} - \theta^{(1)}} = \frac{0.6265 - 0.6243}{0.6243 - 0.6082} = 0.1366$$

Whereas the real $M_e^* = DM(\theta^*)$ is 0.1328 shown in the fourth column of table 1.1 at $t = 4$.

4. Variants of EM algorithm

The main purpose of EM algorithm (GEM algorithm) is to maximize the log-likelihood $L(\Theta) = \log(g(Y \mid \Theta))$ with observed data (incomplete data) Y by maximizing the condition expectation $Q(\Theta' \mid \Theta)$. Such $Q(\Theta' \mid \Theta)$ is defined fixedly in E-step. Therefore, most variants of EM algorithm focus on how to maximize $Q(\Theta' \mid \Theta)$ in M-step more effectively so that EM is faster or more accurate.

4.1. EM algorithm with prior probability

DLR (Dempster, Laird, & Rubin, 1977, pp. 6, 11) mentioned that the convergence rate $DM(\Theta^*)$ specified by equation 3.11 can be improved by adding a prior probability $\pi(\Theta)$ in conjugation with $f(X \mid \Theta)$, $g(Y \mid \Theta)$ or $k(X \mid Y, \Theta)$ according to maximum a posteriori probability (MAP)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2018 doi:10.20944/preprints201802.0131.v1

method (Wikipedia, Maximum a posteriori estimation, 2017). For example, if $\pi(\Theta)$ in conjugation with $g(Y \mid \Theta)$ then, the posterior probability $\pi(\Theta \mid Y)$ is:

$$\pi(\Theta|Y) = \frac{g(Y|\Theta)\pi(\Theta)}{\int_\Theta g(Y|\Theta)\pi(\Theta)\mathrm{d}\Theta}$$

Because $\int_\Theta g(Y|\Theta)\pi(\Theta)\mathrm{d}\Theta$ is constant with regard to Θ, the optimal likelihood-maximization estimate Θ^* is a maximizer of $g(Y \mid \Theta)\pi(\Theta)$. When $\pi(\Theta)$ is conjugate prior of the posterior probability $\pi(\Theta \mid X)$ (or $\pi(\Theta \mid Y)$), both $\pi(\Theta)$ and $\pi(\Theta \mid X)$ (or $\pi(\Theta \mid Y)$) have the same distributions (Wikipedia, Conjugate prior, 2018); for example, if $\pi(\Theta)$ is distributed normally, $\pi(\Theta \mid X)$ (or $\pi(\Theta \mid Y)$) is also distributed normally.

For GEM algorithm, the log-likelihood function associated MAP method is $\mathcal{L}(\Theta)$ specified by equation 4.1.1 with note that $\pi(\Theta)$ is non-convex function.

$$\mathcal{L}(\Theta) = \log\big(g(Y|\Theta)\pi(\Theta)\big) = L(\Theta) + \log\big(\pi(\Theta)\big) \tag{4.1.1}$$

It implies from equation 3.2 that

$$Q(\Theta'|\Theta) + \log\big(\pi(\Theta')\big) = L(\Theta') + \log\big(\pi(\Theta')\big) + H(\Theta'|\Theta) = \mathcal{L}(\Theta') + H(\Theta'|\Theta)$$

Let,

$$Q_+(\Theta'|\Theta) = Q(\Theta'|\Theta) + \log\big(\pi(\Theta')\big) \tag{4.1.2}$$

GEM algorithm now aims to maximize $Q_+(\Theta' \mid \Theta)$ instead of maximizing $Q(\Theta' \mid \Theta)$. The proof of convergence for $Q_+(\Theta' \mid \Theta)$ is not changed in manner but determining the convergence matrix M_e for $Q_+(\Theta' \mid \Theta)$ is necessary. Because $H(\Theta' \mid \Theta)$ is kept intact whereas $Q(\Theta' \mid \Theta)$ is replaced by $Q_+(\Theta' \mid \Theta)$, we expect that the convergence rate m^* specified by equation 3.23 is smaller so that the convergence speed s^* is increased and so GEM algorithm is improved with regard to $Q_+(\Theta' \mid \Theta)$. Equation 4.1.3 specifies $DM(\Theta^*)$ for $Q_+(\Theta' \mid \Theta)$.

$$DM(\Theta^*) = D^{20}H(\Theta^*|\Theta^*)\big(D^{20}Q_+(\Theta^*|\Theta^*)\big)^{-1} \tag{4.1.3}$$

Where $Q_+(\Theta' \mid \Theta)$ is specified by equation 3.32 and $D^{20}Q_+(\Theta' \mid \Theta)$ is specified by equation 4.1.4.

$$D^{20}Q_+(\Theta'|\Theta) = D^{20}Q(\Theta'|\Theta) + D^{20}L\big(\pi(\Theta')\big) \tag{4.1.4}$$

Where,

$$L\big(\pi(\Theta')\big) = \log\big(\pi(\Theta')\big)$$

Suppose all partial derivatives of $Q(\Theta' \mid \Theta)$ and $\pi(\Theta')$ are continuous so that $D^{20}Q(\Theta^* \mid \Theta^*)$ and $D^{20}L(\pi(\Theta^*))$ are symmetric matrices according to Schwarz's theorem (Wikipedia, Symmetry of second derivatives, 2018). Thus, $D^{20}Q(\Theta^* \mid \Theta^*)$ and $D^{20}L(\pi(\Theta^*))$ are commutative:

$$D^{20}Q(\Theta^* \mid \Theta^*)D^{20}L(\pi(\Theta^*)) = D^{20}L(\pi(\Theta^*))D^{20}Q(\Theta^* \mid \Theta^*)$$

Suppose both $D^{20}Q(\Theta^* \mid \Theta^*)$ and $D^{20}L(\pi(\Theta^*))$ are diagonalizable then, they are simultaneously diagonalizable (Wikipedia, Commuting matrices, 2017). Hence there is a (orthogonal) eigenvector matrix V such that (Wikipedia, Diagonalizable matrix, 2017) (StackExchange, 2013):

$$D^{20}Q(\Theta^*|\Theta^*) = VQ_e^*V^{-1}$$

$$D^{20}L\big(\pi(\Theta^*)\big) = V\Pi_e^*V^{-1}$$

Where Q_e^* and Π_e^* are eigenvalue matrices of $D^{20}Q(\Theta^* \mid \Theta^*)$ and $D^{20}L(\pi(\Theta^*))$, respectively. Note Q_e^* and its eigenvalues are mentioned in equation 3.20. Because $\pi(\Theta^*)$ is non-convex function, eigenvalues $\pi_1^*, \pi_2^*, \ldots, \pi_r^*$ of Π_e^* are non-positive.

$$\Pi_e^* = \begin{pmatrix} \pi_1^* & 0 & \cdots & 0 \\ 0 & \pi_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_r^* \end{pmatrix}$$

From equation 4.1.2, $D^{20}Q_+(\Theta^* \mid \Theta^*)$ is decomposed as below:

$$D^{20}Q_+(\Theta^*|\Theta^*) = D^{20}Q(\Theta^*|\Theta^*) + D^{20}L\big(\pi(\Theta^*)\big) = VQ_e^*V^{-1} + V\Pi_e^*V^{-1} = V(Q_e^* + \Pi_e^*)V^{-1}$$

So eigenvalue matrix of $D^{20}Q_+(\Theta^* \mid \Theta^*)$ is $(Q_e^* + \Pi_e^*)$ and eigenvalues of $D^{20}Q_+(\Theta^* \mid \Theta^*)$ are $q_i^* + \pi_i^*$, as follows:

$$Q_e^* + \Pi_e^* = \begin{pmatrix} q_1^* + \pi_1^* & 0 & \cdots & 0 \\ 0 & q_2^* + \pi_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_r^* + \pi_r^* \end{pmatrix}$$

According to equation 3.19, the eigenvalue matrix of $D^{20}H(\Theta^* \mid \Theta^*)$ is H_e^* fixed as follows:

$$H_e^* = \begin{pmatrix} h_1^* & 0 & \cdots & 0 \\ 0 & h_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_r^* \end{pmatrix}$$

Due to $DM(\Theta^*) = D^{20}H(\Theta^* \mid \Theta^*)D^{20}Q_+(\Theta^* \mid \Theta^*)$, equation 3.21 is re-calculated:

$$DM(\Theta^*) = (UH_e^*U^{-1})(U(Q_e^* + \Pi_e^*)U^{-1})^{-1} = UH_e^*U^{-1}U(Q_e^* + \Pi_e^*)^{-1}U^{-1}$$
$$= U(H_e^*(Q_e^* + \Pi_e^*)^{-1})U^{-1}$$

As a result, the convergence matrix M_e^* which is eigenvalue matrix of $DM(\Theta^*)$ is re-calculated by equation 4.1.5.

$$M_e^* = H_e^*(Q_e^* + \Pi_e^*)^{-1} = \begin{pmatrix} m_1^* = \frac{h_1^*}{q_1^* + \pi_1^*} & 0 & \cdots & 0 \\ 0 & m_2^* = \frac{h_2^*}{q_2^* + \pi_2^*} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_r^* = \frac{h_r^*}{q_r^* + \pi_r^*} \end{pmatrix} \quad (4.1.5)$$

The convergence rate m^* of GEM is re-defined by equation 4.1.6.

$$m^* = \max_{m_i^*}\{m_1^*, m_2^*, \dots, m_r^*\} \text{ where } m_i^* = \frac{h_i^*}{q_i^* + \pi_i^*} \quad (4.1.6)$$

Because all h_i^*, q_i^*, and π_i^* are non-positive, we have:

$$\frac{h_i^*}{q_i^* + \pi_i^*} \leq \frac{h_i^*}{q_i^*}, \forall i$$

Therefore, by comparing equation 4.1.6 and equation 3.24, we conclude that m^* is smaller with regard to $Q_+(\Theta' \mid \Theta)$. In other words, the convergence rate is improved with support of prior probability $\pi(\Theta)$. In literature of EM, the combination of GEM and MAP with support of $\pi(\Theta)$ results out a so-called MAP-GEM algorithm.

4.2. EM algorithm with Newton-Raphson method

In the M-step of GEM algorithm, the next estimate $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$, which means that $\Theta^{(t+1)}$ is a solution of equation $D^{10}Q(\Theta \mid \Theta^{(t)}) = \mathbf{0}^T$ where $D^{10}Q(\Theta \mid \Theta^{(t)})$ is the first-order derivative of $Q(\Theta \mid \Theta^{(t)})$ with regard to variable Θ. Note, $Q(\Theta \mid \Theta^{(t)})$ is concave with regard to Θ. Newton-Raphson method (McLachlan & Krishnan, 1997, p. 29) is applied into solving the equation $D^{10}Q(\Theta \mid \Theta^{(t)}) = \mathbf{0}^T$. As a result, M-step is replaced a so-called Newton step (N-step).

N-step starts with an arbitrary value Θ_0 as a solution candidate and also goes through many iterations. Suppose the current parameter is Θ_i, the next value Θ_{i+1} is calculated based on equation 4.2.1.

$$\Theta_{i+1} = \Theta_i - \left(D^{20}Q\left(\Theta_i \middle| \Theta^{(t)}\right)\right)^{-1}\left(D^{10}Q\left(\Theta_i \middle| \Theta^{(t)}\right)\right)^T \quad (4.2.1)$$

N-step converges after some i^{th} iteration. At that time, Θ_{i+1} is solution of equation $D^{10}Q(\Theta \mid \Theta^{(t)}) = 0$ if $\Theta_{i+1}=\Theta_i$. So the next parameter of GEM is $\Theta^{(t+1)} = \Theta_{i+1}$. The equation 4.2.1 is Newton-Raphson process. Recall that $D^{10}Q(\Theta \mid \Theta^{(t)})$ is gradient vector and $D^{20}Q(\Theta \mid \Theta^{(t)})$ is Hessian matrix. Following is a proof of equation 4.2.1.

According to first-order Taylor series expansion of $D^{10}Q(\Theta \mid \Theta^{(t)})$ at $\Theta = \Theta_i$ with very small residual, we have:

$$D^{10}Q\left(\Theta\middle|\Theta^{(t)}\right) = D^{10}Q\left(\Theta_i\middle|\Theta^{(t)}\right) + (\Theta - \Theta_i)^T\left(D^{20}Q\left(\Theta\middle|\Theta^{(t)}\right)\right)^T$$

Suppose all partial derivatives of $Q(\Theta' \mid \Theta)$ are continuous so that $D^{20}Q(\Theta \mid \Theta^{(t)})$ is symmetric matrix according to Schwarz's theorem (Wikipedia, Symmetry of second derivatives, 2018).

$$D^{20}Q(\Theta \mid \Theta^{(t)}) = (D^{20}Q(\Theta \mid \Theta^{(t)}))^T$$

So we have:

$$D^{10}Q\left(\Theta\middle|\Theta^{(t)}\right) = D^{10}Q\left(\Theta_i\middle|\Theta^{(t)}\right) + (\Theta - \Theta_i)^T D^{20}Q\left(\Theta_i\middle|\Theta^{(t)}\right)$$

Let $\Theta = \Theta_{i+1}$ and we expect that $D^{10}Q(\Theta_{i+1} \mid \Theta^{(t)}) = \mathbf{0}^T$ so that Θ_{i+1} is a solution.

$$\mathbf{0}^T = D^{10}Q\left(\Theta_{i+1}\middle|\Theta^{(t)}\right) = D^{10}Q\left(\Theta_i\middle|\Theta^{(t)}\right) + (\Theta_{i+1} - \Theta_i)^T D^{20}Q\left(\Theta_i\middle|\Theta^{(t)}\right)$$

It implies:

$$(\Theta_{i+1})^T = (\Theta_i)^T - D^{10}Q\left(\Theta_i\middle|\Theta^{(t)}\right)\left(D^{20}Q\left(\Theta_i\middle|\Theta^{(t)}\right)\right)^{-1}$$

This means:

$$\Theta_{i+1} = \Theta_i - \left(D^{20}Q\left(\Theta_i\middle|\Theta^{(t)}\right)\right)^{-1}\left(D^{10}Q\left(\Theta_i\middle|\Theta^{(t)}\right)\right)^T \blacksquare$$

Rai and Matthews (Rai & Matthews, 1993) proposed a so-called EM1 algorithm in which Newton-Raphson process is reduced into one iteration, as seen in table 4.2.1 (Rai & Matthews, 1993, pp. 587-588). Rai and Matthews assumed that $f(x)$ belongs to exponential family but their EM1 algorithm is really a variant of GEM in general. In other words, there is no requirement of exponential family for EM1 but $Q(\Theta \mid \Theta^{(t)})$ must be concave ($D^{20}Q(\Theta \mid \Theta^{(t)})$ negative definite).

E-step: The expectation $Q(\Theta \mid \Theta^{(t)})$ is calculated based on current $\Theta^{(t)}$, according to equation 2.4. *M-step*: The next parameter $\Theta^{(t+1)}$ is: $\Theta^{(t+1)} = \Theta^{(t)} - \left(D^{20}Q\left(\Theta^{(t)}\middle\|\Theta^{(t)}\right)\right)^{-1}\left(D^{10}Q\left(\Theta^{(t)}\middle\|\Theta^{(t)}\right)\right)^T$ (4.2.2)

Table 4.2.1. E-step and M-step of EM1 algorithm

Rai and Matthews proved convergence of EM1 algorithm by their proposal of equation 4.2.2. By second-order Taylor series expansion of $Q(\Theta \mid \Theta^{(t)})$ at $\Theta = \Theta^{(t+1)}$ with very small residual, we have:

$$Q\left(\Theta\middle|\Theta^{(t)}\right) = Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) + D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta - \Theta^{(t+1)}\right) + \frac{1}{2}\left(\Theta - \Theta^{(t+1)}\right)^T D^{20}\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta - \Theta^{(t+1)}\right)$$

Let $\Theta = \Theta^{(t)}$, we have:

$$Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) = D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta^{(t+1)} - \Theta^{(t)}\right) - \frac{1}{2}\left(\Theta^{(t+1)} - \Theta^{(t)}\right)^T D^{20}\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta^{(t+1)} - \Theta^{(t)}\right)$$

By substituting equation 4.2.2 for $Q(\Theta^{(t+1)} \mid \Theta^{(t)}) - Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ with note that $D^{20}Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is symmetric matrix, we have:

$$Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) = -D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) * \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1} * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T - \frac{1}{2}D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) * \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1} * D^{20}\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) * \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1} * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2018 doi:10.20944/preprints201802.0131.v1

$$\left(\text{Due to } \left(\left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1}\right)^T = \left(\left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T\right)^{-1} = \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1}\right)$$

So we have:

$$Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)$$
$$= -\frac{3}{2}D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) * \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1} * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T$$

Because $D^{20}Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is negative definite, we have:

$$Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) > 0 \blacksquare$$

Hence, EM1 algorithm satisfies the definition 1 of GEM algorithm and so convergence of EM1 is asserted according corollary 3.

Rai and Matthews made experiment on their EM1 algorithm (Rai & Matthews, 1993, p. 590). As a result, EM1 algorithm saved a lot of computations in M-step. In fact, by comparing GEM (table 2.2) and EM1 (table 4.2.1), EM1 does not maximize $Q(\Theta \mid \Theta^{(t)})$ in each iteration as GEM does but $Q(\Theta \mid \Theta^{(t)})$ will be maximized in the last iteration when EM1 converges. EM1 gains this excellent and interesting result because of Newton-Raphson process specified by equation 4.2.2.

Because equations 3.12 and 3.13 are not changed with regard to EM1, the convergence matrix of EM1 is not changed.

$$M_e = H_e Q_e^{-1}$$

Therefore, EM1 does not improve convergence rate in theory as MAP-GEM algorithm does but EM1 algorithm really speeds up GEM process in practice because it saves computational cost in M-step.

In equation 4.2.2, the second-order derivative $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ is re-computed at every iteration for each Θ(t). If $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ is complicated, it can be fixed by $D^{20}Q(\Theta^{(0)} \mid \Theta^{(0)})$ over all iterations where $\Theta^{(0)}$ is arbitrarily initialized for EM process so as to save computational cost. In other words, equation 4.2.2 is replaced by equation 4.2.3 (Ta, 2014).

$$\Theta^{(t+1)} = \Theta^{(t)} - \left(D^{20}Q\left(\Theta^{(0)}\middle|\Theta^{(0)}\right)\right)^{-1}\left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T \quad (4.2.3)$$

In equation 4.2.3, only $D^{10}Q(\Theta^{(t)} \mid \Theta^{(t)})$ is re-computed at every iteration whereas $D^{20}Q(\Theta^{(0)} \mid \Theta^{(0)})$ is fixed. Equation 4.2.3 implies a pseudo Newton-Raphson process which still converges to Θ^* but it is slower than Newton-Raphson process specified by equation 4.2.2 (Ta, 2014).

Newton-Raphson process specified by equation 4.2.2 has second-order convergence. I propose to use equation 4.2.4 for speeding up EM1 algorithm. In other words, equation 4.2.2 is replaced by equation 4.2.4 (Ta, 2014), in which Newton-Raphson process is improved with third-order convergence. Note, equation 4.2.4 is common in literature of Newton-Raphson process.

$$\Theta^{(t+1)} = \Theta^{(t)} - \left(D^{20}Q\left(\Phi^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1}\left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T$$

Where, (4.2.4)

$$\Phi^{(t)} = \Theta^{(t)} - \frac{1}{2}\left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1}\left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T$$

Similar to equation 4.2.2, equation 4.2.4 satisfies definition 1, theorem 1, corollary 1, corollary 2, and theorem 2. The convergence of equation 4.2.4 is also asserted. Following is a proof of equation 4.2.4 by Ta (Ta, 2014).

Without loss of generality, suppose Θ is scalar such that $\Theta = \theta$, let

$$q(\theta) = D^{10}Q\left(\theta\middle|\theta^{(t)}\right)$$

Let $r(\theta)$ represents improved Newton-Raphson process.

$$\eta(\theta) = \theta - \frac{q(\theta)}{q'\left(\theta + \omega(\theta)q(\theta)\right)}$$

Suppose $\omega(\theta)$ has first derivative and we will find $\omega(\theta)$. The first-order of $\eta(\theta)$ is (Ta, 2014):

$$\eta'(\theta) = 1 - \frac{q'(\theta)}{q'\big(\theta + \omega(\theta)q(\theta)\big)} + \frac{q(\theta)q''\big(\theta + \omega(\theta)q(\theta)\big)\big(1 + \omega'(\theta)q(\theta) + \omega(\theta)q'(\theta)\big)}{\Big(q'\big(\theta + \omega(\theta)q(\theta)\big)\Big)^2}$$

The second-order of $\eta(\theta)$ is (Ta, 2014):

$$\eta''(\theta) = -\frac{q''(\theta)}{q'\big(\theta + \omega(\theta)q(\theta)\big)} + \frac{2q'(\theta)q''\big(\theta + \omega(\theta)q(\theta)\big)\big(1 + \omega'(\theta)q(\theta) + \omega(\theta)q'(\theta)\big)}{\Big(q'\big(\theta + \omega(\theta)q(\theta)\big)\Big)^2} - \frac{2q(\theta)\Big(q''\big(\theta + \omega(\theta)q(\theta)\big)\Big)^2\big(1 + \omega'(\theta)q(\theta) + \omega(\theta)q'(\theta)\big)^2}{\Big(q'\big(\theta + \omega(\theta)q(\theta)\big)\Big)^3} + \frac{q(\theta)q'''\big(\theta + \omega(\theta)q(\theta)\big)\big(1 + \omega'(\theta)q(\theta) + \omega(\theta)q'(\theta)\big)^2}{\Big(q'\big(\theta + \omega(\theta)q(\theta)\big)\Big)^2} + \frac{\big(q(\theta)\big)^2 q''\big(\theta + \omega(\theta)q(\theta)\big)\omega''(\theta)}{\Big(q'\big(\theta + \omega(\theta)q(\theta)\big)\Big)^2} + \frac{q(\theta)q''\big(\theta + \omega(\theta)q(\theta)\big)\big(2\omega'(\theta)q'(\theta) + \omega(\theta)q''(\theta)\big)}{\Big(q'\big(\theta + \omega(\theta)q(\theta)\big)\Big)^2}$$

If $\bar{\theta}$ is solution of equation $q(\theta) = 0$, we have (Ta, 2014):

$$q(\bar{\theta}) = 0$$
$$\eta(\bar{\theta}) = \bar{\theta}$$
$$\eta'(\bar{\theta}) = 0$$
$$\eta''(\bar{\theta}) = \frac{q''(\bar{\theta})}{q'(\bar{\theta})}\big(1 + 2\omega(\bar{\theta})q'(\bar{\theta})\big)$$

We expect $\eta''(\bar{\theta}) = 0$ and so we select (Ta, 2014):

$$\omega(\theta) = -\frac{q(\theta)}{2q'(\theta)}, \forall\theta$$

It means that the Newton-Raphson process is improved as follows (Ta, 2014):

$$\theta^{(t+1)} = \theta^{(t)} - \frac{q\big(\theta^{(t)}\big)}{q'\left(\theta^{(t)} - \frac{q\big(\theta^{(t)}\big)}{2q'\big(\theta^{(t)}\big)}\right)}$$

This means:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{D^{10}Q\big(\theta \big| \theta^{(t)}\big)}{D^{20}Q\left(\theta^{(t)} - \frac{D^{10}Q\big(\theta \big| \theta^{(t)}\big)}{2D^{20}Q\big(\theta \big| \theta^{(t)}\big)} \middle| \theta^{(t)}\right)}$$

As a result, equation 4.2.4 is a generality of the equation above when Θ is vector.

4.3. EM algorithm with Aitken acceleration

According to Lansky and Casella (Lansky & Casella, 1992), GEM converges faster by combination of GEM and Aitken acceleration. Without loss of generality, suppose Θ is scalar such that $\Theta = \theta$, the sequence $\{\theta\}_{t=1}^{+\infty} = \theta^{(1)}, \theta^{(2)}, \dots, \theta^{(t)}, \dots$ is monotonous. From equation 3.28

$$DM(\theta^*) = \lim_{t\to+\infty} \frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*}$$

We have the following approximate with *t* large enough (Lambers, 2009, p. 1):

$$\frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*} \approx \frac{\theta^{(t+2)} - \theta^*}{\theta^{(t+1)} - \theta^*}$$

We establish the following equation from the above approximation, as follows (Lambers, 2009, p. 1):

$$\frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*} \approx \frac{\theta^{(t+2)} - \theta^*}{\theta^{(t+1)} - \theta^*}$$

$$\Rightarrow \left(\theta^{(t+1)} - \theta^*\right)^2 \approx \left(\theta^{(t+2)} - \theta^*\right)\left(\theta^{(t)} - \theta^*\right)$$

$$\Rightarrow \left(\theta^{(t+1)}\right)^2 - 2\theta^{(t+1)}\theta^* \approx \theta^{(t+2)}\theta^{(t)} - \theta^{(t+2)}\theta^* - \theta^{(t)}\theta^*$$

$$\Rightarrow \left(\theta^{(t+2)} - 2\theta^{(t+1)} + \theta^{(t)}\right)\theta^* \approx \theta^{(t)}\left(\theta^{(t+2)} - 2\theta^{(t+1)} + \theta^{(t)}\right) - \left(\theta^{(t+1)} - \theta^{(t)}\right)^2$$

Hence, θ^* is approximated by (Lambers, 2009, p. 1)

$$\theta^* \approx \theta^{(t)} - \frac{\left(\theta^{(t+1)} - \theta^{(t)}\right)^2}{\theta^{(t+2)} - 2\theta^{(t+1)} + \theta^{(t)}}$$

We construct Aitken sequence $\left\{\hat{\theta}\right\}_{t=1}^{+\infty} = \hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots, \hat{\theta}^{(t)}, \dots$ such that (Wikipedia, Aitken's delta-squared process, 2017)

$$\hat{\theta}^{(t)} = \theta^{(t)} - \frac{\left(\theta^{(t+1)} - \theta^{(t)}\right)^2}{\theta^{(t+2)} - 2\theta^{(t+1)} + \theta^{(t)}} = \theta^{(t)} - \frac{\left(\Delta\theta^{(t)}\right)^2}{\Delta^2\theta^{(t)}} \quad (4.3.1)$$

Where Δ is forward difference operator,

$$\Delta\theta^{(t)} = \theta^{(t+1)} - \theta^{(t)}$$

And

$$\Delta^2\theta^{(t)} = \Delta\left(\Delta\theta^{(t)}\right) = \Delta\left(\theta^{(t+1)} - \theta^{(t)}\right) = \Delta\theta^{(t+1)} - \Delta\theta^{(t)}$$
$$= \left(\theta^{(t+2)} - \theta^{(t+1)}\right) - \left(\theta^{(t+1)} - \theta^{(t)}\right) = \theta^{(t+2)} - 2\theta^{(t+1)} + \theta^{(t)}$$

When Θ is vector as $\Theta = (\theta_1, \theta_2, \dots, \theta_r)^T$, Aitken sequence $\left\{\hat{\Theta}\right\}_{t=1}^{+\infty} = \hat{\Theta}^{(1)}, \hat{\Theta}^{(2)}, \dots, \hat{\Theta}^{(t)}, \dots$ is defined by applying equation 4.3.1 into its components θ_i (s) according to equation 4.3.2:

$$\hat{\theta}_i^{(t)} = \theta_i^{(t)} - \frac{\left(\Delta\theta_i^{(t)}\right)^2}{\Delta^2\theta_i^{(t)}}, \forall i = 1, 2, \dots, r \quad (4.3.2)$$

Where,

$$\Delta\theta_i^{(t)} = \theta_i^{(t+1)} - \theta_i^{(t)}$$
$$\Delta^2\theta^{(t)} = \theta_i^{(t+2)} - 2\theta_i^{(t+1)} + \theta_i^{(t)}$$

According theorem of Aitken acceleration, Aitken sequence $\left\{\hat{\Theta}\right\}_{t=1}^{+\infty}$ approaches Θ^* faster than the sequence $\{\Theta\}_{t=1}^{+\infty} = \Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(t)}, \dots$ with note that the sequence $\{\Theta\}_{t=1}^{+\infty}$ is instance of GEM.

$$\lim_{t\to+\infty} \frac{\hat{\theta}_i^{(t)} - \theta^*}{\theta_i^{(t)} - \theta^*} = 0$$

Essentially, the combination of GEM and Aitken acceleration is to replace the sequence $\{\Theta\}_{t=1}^{+\infty}$ by Aitken sequence $\{\widehat{\Theta}\}_{t=1}^{+\infty}$ as seen in table 4.3.1.

E-step: The expectation $Q(\Theta \mid \Theta^{(t)})$ is calculated based on current $\Theta^{(t)}$, according to equation 2.4. Note that $t = 1, 2, 3,\ldots$ and $\Theta^{(0)} = \Theta^{(1)}$. *M-step*: Let $\Theta^{(t+1)} = (\theta_1^{(t+1)}, \theta_2^{(t)},\ldots, \theta_r^{(t+1)})^T$ be a maximizer of $Q(\Theta \mid \Theta^{(t)})$. Note $\Theta^{(t+1)}$ will become current parameter at the next iteration ((t+1)$^{\text{th}}$ iteration). Aitken parameter $\widehat{\Theta}^{(t-1)} = \left(\hat{\theta}_1^{(t-1)}, \hat{\theta}_2^{(t-1)}, \ldots, \hat{\theta}_r^{(t-1)}\right)^T$ is calculated according to equation 4.3.2. $$\hat{\theta}_i^{(t-1)} = \theta_i^{(t-1)} - \frac{\left(\Delta\theta_i^{(t-1)}\right)^2}{\Delta^2\theta_i^{(t-1)}}$$ If $\widehat{\Theta}^{(t-1)} = \widehat{\Theta}^{(t-2)}$ then, the algorithm stops and we have $\widehat{\Theta}^{(t-1)} = \widehat{\Theta}^{(t-2)} = \Theta^*$.

Table 4.3.1. E-step and M-step of GEM algorithm combined with Aitken acceleration

Because Aitken sequence $\{\widehat{\Theta}\}_{t=1}^{+\infty}$ converges to Θ^* faster than the sequence $\{\Theta\}_{t=1}^{+\infty}$ does, the convergence of GEM is improved with support of Aitken acceleration method.

In equation 4.3.2, parametric components θ_i (s) converges separately. Guo, Li, and Xu (Guo, Li, & Xu, 2017) assumed such components converges together with the same rate. So they replaced equation 4.3.2 by equation 4.3.3 (Guo, Li, & Xu, 2017, p. 176) for Aitken sequence $\{\widehat{\Theta}\}_{t=1}^{+\infty}$.

$$\widehat{\Theta}^{(t)} = \Theta^{(t)} - \frac{\left|\Delta\Theta^{(t)}\right|^2}{\left|\Delta^2\Theta^{(t)}\right|}\Delta^2\Theta^{(t)} \tag{4.3.3}$$

4.4. ECM algorithm

Because M-step of GEM is complicated, Meng and Rubin (Meng & Rubin, 1993) proposed a so-called Expectation Conditional Expectation (ECM) algorithm in which M-step is replaced by several computationally simpler Conditional Maximization (CM) steps. Each CM-step maximizes $Q(\Theta \mid \Theta^{(t)})$ on given constraint. ECM is very useful in the case that maximization of $Q(\Theta \mid \Theta^{(t)})$ with constraints is simpler than maximization of $Q(\Theta \mid \Theta^{(t)})$ without constraints as usual.

Suppose the parameter Θ is partitioned into S sub-parameters $\Theta = \{\Theta_1, \Theta_2,\ldots, \Theta_S\}$ and there are S pre-selected vector function $g_s(\Theta)$:

$$G = \{g_s(\Theta); s = 1,2, \ldots, S\} \tag{4.4.1}$$

Each function $g_s(\Theta)$ represents a constraint. Support there is a sufficient enough number of derivatives of each $g_s(\Theta)$. In ECM algorithm (Meng & Rubin, 1993, p. 268), M-step is replaced by a sequence of CM-steps. Each CM-step maximizes $Q(\Theta \mid \Theta^{(t)})$ over Θ but with some function $g_s(\Theta)$ fixed at its previous value. Concretely, there are S CM-steps and every s^{th} CM-step finds $\Theta^{(t+s/S)}$ that maximizes $Q(\Theta \mid \Theta^{(t)})$ over Θ subject to the constraint $g_s(\Theta) = g_s(\Theta^{(t+(s-1)/S)})$. The next parameter $\Theta^{(t+1)}$ is the output of the final CM-step such that $\Theta^{(t+1)} = \Theta^{(t+s/S)}$. Table 4.4.1 (Meng & Rubin, 1993, p. 272) shows E-step and CM-steps of ECM algorithm.

E-step: As usual, the expectation $Q(\Theta \mid \Theta^{(t)})$ is calculated based on current $\Theta^{(t)}$ according to equation 2.4. *CM-steps*: There are S CM-steps. In every s^{th} CM step (s =1, 2,…, S), finding

$$\Theta^{\left(t+\frac{s}{S}\right)} = \underset{\Theta}{\operatorname{argmax}}\left\{Q\left(\Theta \middle| \Theta^{(t)}\right) \text{ with subject to } g_s(\Theta) = g_s\left(\Theta^{\left(t+\frac{s-1}{S}\right)}\right)\right\} \quad (4.4.2)$$ The next parameter $\Theta^{(t+1)}$ is the output of the final CM-step (S^{th} CM-step): $$\Theta^{(t+1)} = \Theta^{\left(t+\frac{S}{S}\right)} \quad (4.4.3)$$ Note, $\Theta^{(t+1)}$ will become current parameter at the next iteration ((t+1)th iteration).

Table 4.3.1. E-step and CM-steps of ECM algorithm

ECM algorithm stops at some t^{th} iteration such that $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$. CM-steps depend on how to define pre-selected functions in G. For example, if $g_s(\Theta)$ consists all sub-parameters except Θ_s then, the s^{th} CM-step maximizes $Q(\Theta \mid \Theta^{(t)})$ with regard to Θ_s whereas other sub-parameters are fixed. If $g_s(\Theta)$ consists only Θ_s then, the s^{th} CM-step maximizes $Q(\Theta \mid \Theta^{(t)})$ with regard to all sub-parameters except Θ_s. Note, definition of ECM algorithm is specified by equations 4.4.2 and 4.4.3.

From equations 4.4.2 and 4.4.3, we have:

$$Q\left(\Theta^{(t+1)} \middle| \Theta^{(t)}\right) = Q\left(M\left(\Theta^{(t)}\right) \middle| \Theta^{(t)}\right) \geq Q\left(\Theta^{(t)} \middle| \Theta^{(t)}\right), \forall t$$

Hence, ECM satisfies the definition 1 of GEM algorithm and so convergence of ECM is asserted according corollary 3. However, Meng and Rubin (Meng & Rubin, 1993, pp. 274-276) provided some conditions for convergence of ECM to a maximizer of $L(\Theta)$.

5. Conclusions

The main purpose of EM algorithm (GEM algorithm) is to maximize the log-likelihood $L(\Theta) = \log(g(Y \mid \Theta))$ with observed data (incomplete data) Y. However, it is too difficult to maximize $\log(g(Y \mid \Theta))$ because $g(Y \mid \Theta)$ is not well-defined when $g(Y \mid \Theta)$ is integral of $f(X \mid \Theta)$ given a general mapping function. DLR solved this problem by an iterative process which is an instance of GEM algorithm. The lower-bound (Sean, 2009, pp. 7-8) of $L(\Theta)$ is maximized over many iterations of the iterative process so that $L(\Theta)$ is maximized finally. Such lower-bound is determined indirectly by the condition expectation $Q(\Theta \mid \Theta^{(t)})$ so that maximizing $Q(\Theta \mid \Theta^{(t)})$ is the same to maximizing the lower bound. Suppose $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ at t^{th} iteration, which is also a maximizer of the lower bound at t^{th} iteration. The lower bound is increased after every iteration. As a result, the maximizer Θ^* of the final lower-bound after many iterations will be expected as a maximizer of $L(\Theta)$ in final.

For more explanations, let $lb(\Theta \mid \Theta^{(t)})$ be lower bound of $L(\Theta)$ at the t^{th} iteration (Sean, 2009, p. 7). From equation 3.2, we have:

$$lb(\Theta \mid \Theta^{(t)}) = Q(\Theta \mid \Theta^{(t)}) - H(\Theta^{(t)} \mid \Theta^{(t)})$$

Due to equations 3.2 and 3.3

$$L(\Theta) = Q(\Theta \mid \Theta^{(t)}) - H(\Theta \mid \Theta^{(t)})$$
$$H(\Theta \mid \Theta^{(t)}) \leq H(\Theta^{(t)} \mid \Theta^{(t)})$$

We have:

$$lb(\Theta \mid \Theta^{(t)}) \leq L(\Theta)$$

The lower bound $lb(\Theta \mid \Theta^{(t)})$ has following property (Sean, 2009, p. 7):

$$lb(\Theta^{(t)} \mid \Theta^{(t)}) = Q(\Theta^{(t)} \mid \Theta^{(t)}) - H(\Theta^{(t)} \mid \Theta^{(t)}) = L(\Theta^{(t)})$$

Therefore, the two steps of GEM is interpreted with regard to the lower bound $lb(\Theta \mid \Theta^{(t)})$ as follows:

E-step:

The lower bound $lb(\Theta \mid \Theta^{(t)})$ is re-calculated based on $Q(\Theta \mid \Theta^{(t)})$.

M-step:

The next parameter $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ which is also a maximizer of $lb(\Theta \mid \Theta^{(t)})$ because $H(\Theta^{(t)} \mid \Theta^{(t)})$ is constant. Note that $\Theta^{(t+1)}$ will become current parameter at the next iteration so that the lower bound is increased in the next iteration.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2018 doi:10.20944/preprints201802.0131.v1

Because $Q(\Theta \mid \Theta^{(t)})$ is defined fixedly in E-step, most variants of EM algorithm focus on how to maximize $Q(\Theta' \mid \Theta)$ in M-step more effectively so that EM is faster or more accurate.

References

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. (M. Stone, Ed.) *Journal of the Royal Statistical Society, Series B (Methodological), 39*(1), 1-38.

Dinh, L. T., Pham, D. H., Nguyen, T. X., & Ta, P. D. (2000). *Univariate Analysis - Principles and Practices.* (K. H. Ha, T. V. Ngo, & D. H. Pham, Eds.) Hanoi, Vietnam: Hanoi National University Publisher. Retrieved from http://www.ebook.edu.vn/?page=1.14&view=11156

Guo, X., Li, Q.-y., & Xu, W.-l. (2017, February). Acceleration of the EM Algorithm Using the Vector Aitken Method and Its Steffensen Form. *Acta Mathematicae Applicatae Sinica, 33*(1), 175-182. doi:10.1007/s10255-017-0648-3

Hardle, W., & Simar, L. (2013). *Applied Multivariate Statistical Analysis.* Berlin, Germany: Research Data Center, School of Business and Economics, Humboldt University.

Jebara, T. (2015). *The Exponential Family of Distributions.* Columbia University, Computer Science Department. New York: Columbia Machine Learning Lab. Retrieved April 27, 2016, from http://www.cs.columbia.edu/~jebara/4771/tutorials/lecture12.pdf

Jia, Y.-B. (2013). *Lagrange Multipliers.* Lecture notes on course “Problem Solving Techniques for Applied Computer Science”, Iowa State University of Science and Technology, USA.

Lambers, J. (2009). *Accelerating Convergence.* University of Southern Mississippi, Department of Mathematics. Hattiesburg: University of Southern Mississippi. Retrieved February 15, 2018, from http://www.math.usm.edu/lambers/mat460/fall09/lecture13.pdf

Lansky, D., & Casella, G. (1992). Improving the EM Algorithm. *Computing Science and Statistics*, 420-424. doi:10.1007/978-1-4612-2856-1_67

McLachlan, G., & Krishnan, T. (1997). *The EM Algorithm and Extensions.* New York, NY, USA: John Wiley & Sons. Retrieved from https://books.google.com.vn/books?id=NBawzaWoWa8C

Meng, X.-L., & Rubin, D. B. (1993, June 1). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 80*(2), 267-278. doi:10.2307/2337198

Nguyen, L. (2015). *Matrix Analysis and Calculus* (1st ed.). (C. Evans, Ed.) Hanoi, Vietnam: Lambert Academic Publishing. Retrieved from https://www.shuyuan.sg/store/gb/book/matrix-analysis-and-calculus/isbn/978-3-659-69400-4

Rai, S. N., & Matthews, D. E. (1993, June). Improving the EM Algorithm. (C. A. McGilchrist, Ed.) *Biometrics, 49*(2), 587-591. doi:10.2307/2532570

Rao, R. C. (1955, June). Estimation and tests of significance in factor analysis. *Psychometrika, 20*(2), 93-111. doi:10.1007/BF02288983

Sean, B. (2009). *The Expectation Maximization Algorithm - A short tutorial.* University of Notre Dame, Indiana, Department of Electrical Engineering. Sean Borman's Homepage.

StackExchange. (2013, November 19). *Eigenvalues of the product of 2 symmetric matrices*. (Stack Exchange Network) Retrieved February 9, 2018, from Mathematics StackExchange: https://math.stackexchange.com/questions/573583/eigenvalues-of-the-product-of-2-symmetric-matrices

Ta, P. D. (2014). *Numerical Analysis Lecture Notes.* Vietnam Institute of Mathematics, Numerical Analysis and Scientific Computing. Hanoi: Vietnam Institute of Mathematics. Retrieved 2014

Wikipedia. (2016, March September). *Exponential family*. (Wikimedia Foundation) Retrieved 2015, from Wikipedia website: https://en.wikipedia.org/wiki/Exponential_family

Wikipedia. (2017, May 25). *Aitken's delta-squared process*. (Wikimedia Foundation) Retrieved February 15, 2018, from Wikipedia website: https://en.wikipedia.org/wiki/Aitken%27s_delta-squared_process

Wikipedia. (2017, February 27). *Commuting matrices*. (Wikimedia Foundation) Retrieved February 9, 2018, from Wikipedia website: https://en.wikipedia.org/wiki/Commuting_matrices

Wikipedia. (2017, November 27). *Diagonalizable matrix*. (Wikimedia Foundation) Retrieved February 10, 2018, from Wikipedia website: https://en.wikipedia.org/wiki/Diagonalizable_matrix#Simultaneous_diagonalization

Wikipedia. (2017, March 2). *Maximum a posteriori estimation*. (Wikimedia Foundation) Retrieved April 15, 2017, from Wikipedia website: https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation

Wikipedia. (2018, January 15). *Conjugate prior*. (Wikimedia Foundation) Retrieved February 15, 2018, from Wikipedia website: https://en.wikipedia.org/wiki/Conjugate_prior

Wikipedia. (2018, January 7). *Symmetry of second derivatives*. (Wikimedia Foundation) Retrieved February 10, 2018, from Wikipedia website: https://en.wikipedia.org/wiki/Symmetry_of_second_derivatives

Wu, J. C. (1983, March). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, *11*(1), 95-103. Retrieved from https://projecteuclid.org/euclid.aos/1176346060

Zivot, E. (2009). *Maximum Likelihood Estimation.* Lecture Notes on course "Econometric Theory I: Estimation and Inference (first quarter, second year PhD)", University of Washington, Seattle, Washington, USA.