*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

# 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. Moreover some improvements of EM algorithm are also proposed in the tutorial such as combination of EM and third-order convergence Newton-Raphson process, combination of EM and gradient descent method, and combination of EM and particle swarm optimization (PSO) algorithm.

**Keywords**: expectation maximization, EM, generalized expectation maximization, GEM, EM convergence.

## 1. Introduction

Literature of expectation maximization (EM) algorithm in this 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.

We begin a review of 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 $Z$. Bold and uppercase letters such as $\boldsymbol{X}$ and $\boldsymbol{R}$ denotes algebraic structures such as spaces and fields. By default, vectors are column vectors. For example, given two vectors $X$ and $Y$ and two matrices $A$ and $B$:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{pmatrix} \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{pmatrix}$$

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \quad B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1k} \\ b_{21} & b_{22} & \cdots & b_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & a_{nk} \end{pmatrix}$$

Matrix $A$ is squared if $m = n$. Matrix $\Lambda$ is diagonal if it is squared and its elements outside the main diagonal are zero:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_r \end{pmatrix}$$

Let *I* be identity matrix or unit matrix, as follows:

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

Let superscript "*T*" denote transposition operation for vector and matrix, as follows:

$$X^T = (x_1, x_2, \ldots, x_r)$$

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{r1} \\ a_{12} & a_{22} & \cdots & a_{r2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{rp} \end{pmatrix}$$

Dot product or scalar product of two vectors can be written with transposition operation, as follows:

$$X^T Y = \sum_{i=1}^{r} x_i y_i$$

However, the product $XY^T$ results out a matrix as follows:

$$XY^T = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_r \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_r \\ \vdots & \vdots & \ddots & \vdots \\ x_r y_1 & x_r y_2 & \cdots & x_r y_r \end{pmatrix}$$

The length of module of vector *X* in Euclidean space is:

$$|X| = \sqrt{X^T X} = \sqrt{\sum_{i=1}^{r} x_i^2}$$

The product of two matrices is:

$$AB = C = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mk} \end{pmatrix}$$

$$c_{ij} = \sum_{v=1}^{n} a_{iv} b_{vj}$$

Matrix *A* is symmetric if $a_{ij} = a_{ji}$ for all *i* and *j*. If *A* is symmetric then, $A^T = A$. If both *A* and *B* are symmetric then, they are commutative such that $AB = BA$.

Suppose $f(X)$ is scalar-by-vector function, for example, $f$: $\boldsymbol{R}^r \rightarrow \boldsymbol{R}$ where $\boldsymbol{R}^r$ is *r*-dimensional real vector space. The first-order derivative of $f(X)$ is gradient vector as follows:

$$f'(X) = \nabla f(X) = \frac{\mathrm{d}f(X)}{\mathrm{d}X} = Df(X) = \left( \frac{\partial f(X)}{\partial x_1}, \frac{\partial f(X)}{\partial x_2}, \ldots, \frac{\partial f(X)}{\partial x_r} \right)$$

Where $\frac{\partial f(X)}{\partial x_1}$ is partial derivative of $f$ with regard to $x_i$. So gradient vector is row vector. The second-order derivative of $f(X)$ is called Hessian matrix as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

$$f''(X) = \frac{\mathrm{d}^2 f(X)}{\mathrm{d}X^2} = D^2 f(X) = \begin{pmatrix} \frac{\partial^2 f(X)}{\partial x_1^2} & \frac{\partial^2 f(X)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(X)}{\partial x_1 \partial x_r} \\ \frac{\partial^2 f(X)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(X)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(X)}{\partial x_2 \partial x_r} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(X)}{\partial x_r \partial x_1} & \frac{\partial^2 f(X)}{\partial x_r \partial x_2} & \cdots & \frac{\partial^2 f(X)}{\partial x_r^2} \end{pmatrix}$$

Hessian matrix is squared matrix. Function $f(X)$ is called $n^{\text{th}}$-order analytic function or $n^{\text{th}}$-order smooth function if there is existence and continuity of $k^{\text{th}}$-order derivatives of $f(X)$ where $k = 1, 2, \ldots, n$. Function $f(X)$ is called smooth enough function if $n$ is large enough. According to Schwarz's theorem (Wikipedia, Symmetry of second derivatives, 2018), if $f(X)$ is second-order smooth function then, its Hessian matrix is symmetric.

Now we skim through an introduction of EM algorithm. 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 $\varphi$: $\boldsymbol{X} \to \boldsymbol{Y}$ 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 $\boldsymbol{X}(Y)$.

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

Where $\Theta$ is probabilistic parameter represented as a column vector, $\Theta = (\theta_1, \theta_2, \ldots, \theta_r)^T$ in which each $\theta_i$ is a particular parameter. Note that, $\Theta$ can degrades into a scalar as $\Theta = \theta$. For example, normal distribution has two particular parameters such as mean $\mu$ and variance $\sigma^2$ 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.

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

DLR (Dempster, Laird, & Rubin, 1977, p. 1) considered $\boldsymbol{X}$ as *complete data* and $\boldsymbol{Y}$ as *incomplete data* because the mapping $\varphi$: $\boldsymbol{X} \to \boldsymbol{Y}$ 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 $\boldsymbol{Y}$ and $f(X \mid \Theta)$ in order to determine $g(Y, \Theta)$ and $k(X \mid Y, \Theta)$. Our purpose is to estimate $\Theta$ based on such $\boldsymbol{Y}$ and $f(X \mid \Theta)$. Pioneers in 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):

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

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 $\Omega$ be the convex set such that $\Theta \in \Omega$. If $\Theta$ is restricted only to $\Omega$ then, $f(X \mid \Theta)$ specifies a *regular exponential family*. If $\Theta$ lies in a curved sub-manifold of $\Omega$ 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.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

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

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

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

Where,

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

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

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

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 $Y$.

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

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

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

If $f(X \mid \Theta)$ follows exponential family, the conditional density $k(X \mid Y, \Theta)$ is determined as follows:

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

If $f(X \mid \Theta)$ follows exponential family then, $k(X \mid Y, \Theta)$ also follows exponential family. In fact, we have:

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

The first-order derivative of $\log(a(\Theta \mid Y))$ is:

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

The second-order derivative of $\log(a(\Theta) \mid Y)$ is:

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

Where $V(\tau(X) \mid Y, \Theta)$ is central covariance matrix of $\tau(X)$ given observed $Y$. Table 1.1 is 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.

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

**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:

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

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:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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:

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

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 $\varphi$ 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 $\theta^*$. 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:

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

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

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

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:

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

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

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

So we have:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

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 $\theta^*$:

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

Table 1.2 (Dempster, Laird, & Rubin, 1977, p. 3) lists estimates of $\theta$ 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 $\theta^*$ whereas the fourth column gives the ratio of successive deviations. Later on, we will know that such ratio implies convergence rate.

| $t$ | $\theta^{(t)}$ | $(\theta^{(t)} - \theta^*)$ | $(\theta^{(t+1)} - \theta^*) / (\theta^{(t)} - \theta^*)$ |
|---|---|---|---|
| 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:

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

Note, $\Theta^{(t+1)}$ will become current parameter at the next iteration ($(t+1)^{th}$ 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 $\Theta^*$ 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 $\Theta$, denoted $L(\Theta)$, with respect to observation $Y$.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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 $\Theta$ 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:

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

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

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

The expression $E(\tau(X) \mid \Theta)$ is function of $\Theta$ but $\tau(X)$ is still dependent on $X$. Let $\tau^{(t)}$ be value of $\tau(X)$ at the $t^{\text{th}}$ iteration of EM process, candidate for the best estimate of $\Theta$ 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,

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

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:

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

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

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

It implies:

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

Let $\Theta^{(t)}$ be the current estimate at some $t^{\text{th}}$ 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 $\Theta^*$ is the optimal estimate of EM process because $\Theta^*$ is solution of the equation:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

Thus, we conclude that $\Theta^*$ 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).

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

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

| |
|---|
| *E-step*:<br>The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$, according to equation 2.4.<br>*M-step*:<br>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 $t^{\text{th}}$ 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 by 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.

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

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

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

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

Because $k(X \mid Y, \Theta)$ belongs exponential family, the expectation $E(\tau(X) \mid Y, \Theta)$ is function of $\Theta$, denoted $\tau_\Theta$. 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 $\varphi$: $\boldsymbol{X} \to \boldsymbol{Y}$, 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 $\varphi$ 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 $(\Theta', \Theta)$ in $\Omega$ x $\Omega$,

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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:

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

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

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

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

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

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

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

It implies:

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

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 $\{\Theta^{(t)}\}_{t=1}^{+\infty} = \Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(t)}, \Theta^{(t+1)}, \ldots$ be a sequence of estimates of $\Theta$ 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).

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

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

$$Q(M(\Theta)|\Theta) \geq Q(\Theta|\Theta) ■ \quad (3.5)$$

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 $\Theta$ 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

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

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):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

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) $k(X \mid Y, M(\Theta^*)) = k(X \mid Y, \Theta^*)$ ■

*Proof*. From theorem 1 and the assumption of corollary 1, we have:

$$\begin{cases} L\big(M(\Theta)\big) \geq L(\Theta) \text{ for all } \Theta \in \Omega \\ L(\Theta^*) \geq L(\Theta) \text{ for all } \Theta \in \Omega \end{cases}$$

This implies:

$$\begin{cases} L\big(M(\Theta^*)\big) \geq L(\Theta^*) \\ L\big(M(\Theta^*)\big) \leq L(\Theta^*) \end{cases}$$

As a result,

$$L\big(M(\Theta^*)\big) = L\big(M(\Theta^*)\big)$$

From theorem 1, we also have:

$$Q(M(\Theta^*)|\Theta^*) = Q(\Theta^*|\Theta^*)$$
$$k\big(X\big|Y, M(\Theta^*)\big) = k(X|Y, \Theta^*) \blacksquare$$

**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:

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

*Proof*. From corollary 1 and the assumption of corollary 2, we have:

$$\begin{cases} L\big(M(\Theta^*)\big) = L(\Theta^*) \\ L(\Theta^*) > L(\Theta) \text{ for all } \Theta \in \Omega \text{ and } \Theta \neq \Theta^* \end{cases}$$

If $M(\Theta^*) \neq \Theta^*$, there is a contradiction $L(M(\Theta^*)) = L(\Theta^*) > L(M(\Theta^*))$. Therefore, we have $M(\Theta^*) = \Theta^*$ ■

**Theorem 2** (Dempster, Laird, & Rubin, 1977, p. 7). Suppose $\{\Theta^{(t)}\}_{t=1}^{+\infty}$ is the sequence of estimates resulted from GEM algorithm such that:

(1) The sequence $\{L(\Theta^{(t)})\}_{t=1}^{+\infty} = L(\Theta^{(1)}), L(\Theta^{(2)}), \ldots, L(\Theta^{(t)}), \ldots$ is bounded above, 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 $\{\Theta^{(t)}\}_{t=1}^{+\infty}$ converges to some $\Theta^*$ in the closure of $\Omega$ ■

*Proof*. The sequence $\{L(\Theta^{(t)})\}_{t=1}^{+\infty}$ is non-decreasing according to theorem 1 and is bounded above according to the assumption 1 of theorem 2 and hence, the sequence $\{L(\Theta^{(t)})\}_{t=1}^{+\infty}$ 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 $v \geq 1$:

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

(Due to $L(\Theta^{(t+i-1)}) = Q(\Theta^{(t+i-1)} \mid \Theta^{(t+i-1)}) - Q(\Theta^{(t+i-1)} \mid \Theta^{(t+i-1)})$ according to equation 3.2)

It implies

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

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

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

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 $\left\{\Theta^{(t)}\right\}_{t=1}^{+\infty}$ converges to some $\Theta^{*}$ in the closure of $\Omega$.

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 and so we need mathematical tools of derivative and differential to prove convergence of GEM. We assume that $Q(\Theta' \mid \Theta)$, $L(\Theta)$, $H(\Theta' \mid \Theta)$, and $M(\Theta)$ are smooth enough. As a convention for derivatives of bivariate function, let $D^{ij}$ denote as the derivative (differential) by taking $i^{\text{th}}$-order partial derivative (differential) with regard to first variable and then, taking $j^{\text{th}}$-order partial derivative (differential) 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 $D^{11}Q(\Theta'|\Theta) = \frac{\partial Q\left(\Theta'\middle|\Theta\right)}{\partial\Theta'\partial\Theta}$
- 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 $M(\Theta)$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

**Table 3.1.** Some differentials of $Q(\Theta' | \Theta)$, $H(\Theta' | \Theta)$, $L(\Theta)$, and $M(\Theta)$

When $\Theta'$ and $\Theta$ 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 $\Theta$ in $\Omega$,

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

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

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

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

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:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

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

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

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:

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

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

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

It implies:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

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

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

We have:

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

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

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

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

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

It implies:

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

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

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

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

Due to:

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

We have:

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

**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 $t$, there exists a $\Theta_0^{(t+1)}$ on the line segment joining $\Theta^{(t)}$ and $\Theta^{(t+1)}$ such that

$$Q\big(\Theta^{(t+1)}\big|\Theta^{(t)}\big) - Q\big(\Theta^{(t)}\big|\Theta^{(t)}\big) = -\big(\Theta^{(t+1)} - \Theta^{(t)}\big)^T D^{20}Q\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right)\big(\Theta^{(t+1)} - \Theta^{(t)}\big)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

Furthermore, if $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite, and the sequence $\{L(\Theta^{(t)})\}_{t=1}^{+\infty}$ is bounded above then, the sequence $\{\Theta^{(t)}\}_{t=1}^{+\infty}$ converges to some $\Theta^*$ in the closure of $\Omega$ ■

Note, if $\Theta$ is a scalar parameter, $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ 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 $\Theta = \Theta^{(t+1)}$ to obtain:

$$Q(\Theta|\Theta^{(t)}) = Q(\Theta^{(t+1)}|\Theta^{(t)}) + D^{10}Q(\Theta^{(t+1)}|\Theta^{(t)})(\Theta - \Theta^{(t+1)}) + (\Theta - \Theta^{(t+1)})^T D^{20}Q(\Theta_0^{(t+1)}|\Theta^{(t)})(\Theta - \Theta^{(t+1)})$$
$$= Q(\Theta^{(t+1)}|\Theta^{(t)}) + (\Theta - \Theta^{(t+1)})^T D^{20}Q(\Theta_0^{(t+1)}|\Theta^{(t)})(\Theta - \Theta^{(t+1)}) \quad (\text{due to } D^{10}Q(\Theta^{(t+1)}|\Theta^{(t)}) = \mathbf{0}^T)$$

Where $\Theta_0^{(t+1)}$ is on the line segment joining $\Theta$ and $\Theta^{(t+1)}$. Let $\Theta = \Theta^{(t)}$, we have:

$$Q(\Theta^{(t+1)}|\Theta^{(t)}) - Q(\Theta^{(t)}|\Theta^{(t)}) = -(\Theta^{(t+1)} - \Theta^{(t)})^T D^{20}Q(\Theta_0^{(t+1)}|\Theta^{(t)})(\Theta^{(t+1)} - \Theta^{(t)})$$

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

$$Q(\Theta^{(t+1)}|\Theta^{(t)}) - Q(\Theta^{(t)}|\Theta^{(t)}) = -(\Theta^{(t+1)} - \Theta^{(t)})^T D^{20}Q(\Theta_0^{(t+1)}|\Theta^{(t)})(\Theta^{(t+1)} - \Theta^{(t)}) > 0$$

Whereas,

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

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

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

In other words, the assumption 2 of theorem 2 is satisfied and hence, the sequence $\{\Theta^{(t)}\}_{t=1}^{+\infty}$ converges to some $\Theta^*$ in the closure of $\Omega$ if the sequence $\{L(\Theta^{(t)})\}_{t=1}^{+\infty}$ is bounded above ■

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

(1) $\Theta^{(t)}$ converges to $\Theta^*$ in the closure of $\Omega$.
(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, and

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

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:

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

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

$$D^{10}Q(\Theta_2|\Theta_1) = D^{10}Q(\Theta_2|\Theta^*) + (\Theta_1 - \Theta^*)^T D^{11}Q(\Theta_2|\Theta^*) + R_1(\Theta_1)$$

Where $R_1(\Theta_1)$ is remainder of $D^{10}Q(\Theta_2 \mid \Theta_1)$ with regard to $\Theta_1$ to obtain:

$$D^{10}Q(\Theta_2|\Theta_1) = D^{10}Q(\Theta^*|\Theta^*) + (\Theta_1 - \Theta^*)^T D^{11}Q(\Theta_2|\Theta^*) + R_1(\Theta_1) \quad (3.12)$$

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

$$D^{10}Q(\Theta_2|\Theta_1) = D^{10}Q(\Theta^*|\Theta_1) + (\Theta_2 - \Theta^*)^T D^{20}Q(\Theta^*|\Theta_1) + R_2(\Theta_2)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

Where $R_2(\Theta_2)$ is remainder of $D^{10}Q(\Theta_2 \mid \Theta_1)$ with regard to $\Theta_2$ to obtain:

$$D^{10}Q(\Theta_2|\Theta_1) = D^{10}Q(\Theta^*|\Theta^*) + (\Theta_2 - \Theta^*)^T D^{20}Q(\Theta^*|\Theta_1) + R_2(\Theta_2) \quad (3.13)$$

By summing equation 12 and equation 13, we have:

$$2D^{10}Q(\Theta_2|\Theta_1) = 2D^{10}Q(\Theta^*|\Theta^*) + (\Theta_1 - \Theta^*)^T D^{11}Q(\Theta_2|\Theta^*) + (\Theta_2 - \Theta^*)^T D^{20}Q(\Theta^*|\Theta_1) + R_1(\Theta_1) + R_2(\Theta_2)$$

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,

$$2D^{10}Q(\Theta_2|\Theta_1) = (\Theta_1 - \Theta^*)^T D^{11}Q(\Theta_2|\Theta^*) + (\Theta_2 - \Theta^*)^T D^{20}Q(\Theta^*|\Theta_1) + R_1(\Theta_1) + R_2(\Theta_2)$$

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

$$2D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) = \left(\Theta^{(t)} - \Theta^*\right)^T D^{11}Q\left(\Theta^{(t+1)}\middle|\Theta^*\right) + \left(\Theta^{(t+1)} - \Theta^*\right)^T D^{20}Q\left(\Theta^*\middle|\Theta^{(t)}\right) + R_1\left(\Theta^{(t)}\right) + R_2\left(\Theta^{(t+1)}\right)$$

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

$$\left(\Theta^{(t)} - \Theta^*\right)^T D^{11}Q\left(\Theta^{(t+1)}\middle|\Theta^*\right) + \left(\Theta^{(t+1)} - \Theta^*\right)^T D^{20}Q\left(\Theta^*\middle|\Theta^{(t)}\right) + R_1\left(\Theta^{(t)}\right) + R_2\left(\Theta^{(t+1)}\right) = \mathbf{0}^T$$

It implies:

$$\left(\Theta^{(t+1)} - \Theta^*\right)^T = \left(\Theta^{(t)} - \Theta^*\right)^T \left(-D^{11}Q\left(\Theta^{(t+1)}\middle|\Theta^*\right)\left(D^{20}Q\left(\Theta^*\middle|\Theta^{(t)}\right)\right)^{-1}\right) - \left(R_1\left(\Theta^{(t)}\right) + R_2\left(\Theta^{(t+1)}\right)\right)\left(D^{20}Q\left(\Theta^*\middle|\Theta^{(t)}\right)\right)^{-1}$$

In other words, we have:

$$\left(M\left(\Theta^{(t)}\right) - M(\Theta^*)\right)^T = \left(\Theta^{(t)} - \Theta^*\right)^T \left(-D^{11}Q\left(\Theta^{(t+1)}\middle|\Theta^*\right)\left(D^{20}Q\left(\Theta^*\middle|\Theta^{(t)}\right)\right)^{-1}\right) - \left(R_1\left(\Theta^{(t)}\right) + R_2\left(\Theta^{(t+1)}\right)\right)\left(D^{20}Q\left(\Theta^*\middle|\Theta^{(t)}\right)\right)^{-1}$$

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

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

Due to, when $t$ approaches $+\infty$, we have:

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

And

$$\lim_{t\to+\infty} R_1\left(\Theta^{(t)}\right) = \lim_{\Theta^{(t)}\to\Theta^*} R_1\left(\Theta^{(t)}\right) = 0$$
$$\lim_{t\to+\infty} R_2\left(\Theta^{(t+1)}\right) = \lim_{\Theta^{(t+1)}\to\Theta^*} R_2\left(\Theta^{(t+1)}\right) = 0$$

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

Finally, theorem 4 is proved. By combination of theorems 2, 3, and 4, corollary 3 is criterion of convergence of GEM.

**Corollary 3.** If an algorithm satisfies three following assumptions:

(1) $Q(M(\Theta^{(t)}) \mid \Theta^{(t)}) \geq Q(\Theta^{(t)} \mid \Theta^{(t)})$ for all $t$.
(2) $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite for all $t$.
(3) The sequence $\left\{L\left(\Theta^{(t)}\right)\right\}_{t=1}^{+\infty}$ is bounded above.

Then,

(1) Such algorithm is an GEM and converges to a stationary point $\Theta^*$ of the likelihood function $L(\Theta)$ such that $DL(\Theta^*) = \mathbf{0}^T$.
(2) $D^{20}Q(\Theta^* \mid \Theta^*)$ is negative definite.
(3) Equation 3.11 is retrieved.

Note that $\Theta_0^{(t+1)}$ is on the line segment joining $\Theta^{(t)}$ and $\Theta^{(t+1)}$ according to second-order Taylor series such that

$$\begin{aligned} Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) = {} & Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) + D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta^{(t)} - \Theta^{(t+1)}\right) \\ & + \left(\Theta^{(t)} - \Theta^{(t+1)}\right)^T D^{20}Q\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta^{(t)} - \Theta^{(t+1)}\right)\blacksquare \end{aligned} \tag{3.15}$$

Please see the proof of theorem 3 to understand how to derive equation 3.15. The condition "$D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite" is less strict than the condition "$D^{20}Q(\Theta \mid \Theta^{(t)})$ is negative definite" but more strict than the condition "$D^{20}Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is negative definite". When $t$ approaches $+\infty$, we have such that $\Theta^{(t)} = \Theta_0^{(t+1)} = \Theta^{(t+1)} = \Theta^*$. Hence, $D^{20}Q(\Theta^* \mid \Theta^*)$ is negative definite if $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite. The assumption 1 of corollary 3 implies that the given algorithm is a GEM according to definition 1. Because $Q(\Theta' \mid \Theta)$ is smooth enough, 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. The assumptions 2 and 3 of corollary 3 along with the fact $D^{10}Q(\Theta^{(t+1)} \mid \Theta^{(t)}) = 0$ imply that the sequence $\left\{\Theta^{(t)}\right\}_{t=1}^{+\infty}$ converges to $\Theta^*$ according to theorems 2 and 3, which in turn implies $DL(\Theta^*) = \mathbf{0}^T$, $D^{20}Q(\Theta^* \mid \Theta^*)$ negative definite, and equation 3.11 according to theorem 4.

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

Because $H(\Theta' \mid \Theta)$ and $Q(\Theta' \mid \Theta)$ are smooth enough, $D^{20}H(\Theta^* \mid \Theta^*)$ 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 3.16 and 3.17. 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 $D^{20}Q(\Theta^* \mid \Theta^*)$.

$$H_e^* = \begin{pmatrix} h_1^* & 0 & \cdots & 0 \\ 0 & h_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_r^* \end{pmatrix} \tag{3.16}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

$$Q_e^* = \begin{pmatrix} q_1^* & 0 & \cdots & 0 \\ 0 & q_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_r^* \end{pmatrix} \tag{3.17}$$

From equation 3.11, $DM(\Theta^*)$ is decomposed as seen in equation 3.18.

$$\begin{aligned} DM(\Theta^*) &= (UH_e^*U^{-1})(UQ_e^*U^{-1})^{-1} = UH_e^*U^{-1}U(Q_e^*)^{-1}\Lambda^{-1}U^{-1} \\ &= U(H_e^*(Q_e^*)^{-1})U^{-1} \end{aligned} \tag{3.18}$$

Let $M_e^*$ be eigenvalue matrix of $DM(\Theta^*)$, specified by equation 18. As a convention $M_e^*$ is called convergence matrix.

$$M_e^* = H_e^*(Q_e^*)^{-1} = \begin{pmatrix} m_1^* = \dfrac{h_1^*}{q_1^*} & 0 & \cdots & 0 \\ 0 & m_2^* = \dfrac{h_2^*}{q_2^*} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_r^* = \dfrac{h_r^*}{q_r^*} \end{pmatrix} \tag{3.19}$$

Of course, all $m_i^* = h_i^* / q_i^*$ are eigenvalues of $DM(\Theta^*)$. Corollary 3 implies $q_i^* < 0$ for all $i$. We will prove that $0 \leq m_i^* \leq 1$ for all $i$ by contradiction. Conversely, suppose we *always* have $m_i^* > 1$ or $m_i^* < 0$ for some $i$. When $\Theta$ degrades into scalar as $\Theta = \theta$, we have converse assumption "$DM(\theta^*) = M_e^* > 1$ or $DM(\theta^*) = M_e^* < 0$". Equation 3.11 is re-written as equation 3.20:

$$\begin{aligned} DM(\theta^*) = M_e^* = m^* &= \lim_{t \to +\infty} \frac{M\left(\theta^{(t)}\right) - M(\theta^*)}{\theta^{(t)} - \theta^*} = \lim_{t \to +\infty} \frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*} = \\ &= D^{20}H(\theta^*|\theta^*)\left(D^{20}Q(\theta^*|\theta^*)\right)^{-1} \end{aligned} \tag{3.20}$$

From equation 3.20, the next estimate $\theta^{(t+1)}$ approaches $\theta^*$ when $t \to +\infty$ and so we have:

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

So equation 3.21 is a variant of equation 3.20 (McLachlan & Krishnan, 1997, p. 120).

$$DM(\theta^*) = M_e = m^* = \lim_{t \to +\infty} \frac{\theta^{(t+2)} - \theta^{(t+1)}}{\theta^{(t+1)} - \theta^{(t)}} \tag{3.21}$$

Because the sequence $\left\{L\left(\theta^{(t)}\right)\right\}_{t=1}^{+\infty} = L\left(\theta^{(1)}\right), L\left(\theta^{(2)}\right), \ldots, L\left(\theta^{(t)}\right), \ldots$ is non-decreasing, the sequence $\left\{\theta^{(t)}\right\}_{t=1}^{+\infty} = \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}, \ldots$ 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

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

So we have

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

However, this contradict the converse assumption "$DM(\theta^*) = M_e^* > 1$ or $DM(\theta^*) = M_e^* < 0$". Therefore, we conclude that $0 \leq m_i^* \leq 1$ for all $i$. In general, if $\Theta^*$ is stationary point of GEM then, $D^{20}Q(\Theta^* \mid \Theta^*)$ 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, according to equation 3.22.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

$$\begin{aligned} q_i^* &< 0, \forall i \\ h_i^* &\le 0, \forall i \\ 0 \le m_i^* &\le 1, \forall i \end{aligned} \tag{3.22}$$

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

Suppose $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_r^{(t)})$ at current $t^{\text{th}}$ 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 $\theta_i^*$. 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 3.23. The convergence rate $m^*$ implies lowest speed.

$$m^* = \max_{m_i^*}\{m_1^*, m_2^*, \ldots, m_r^*\} \text{ where } m_1^* = \frac{h_1^*}{q_1^*} \tag{3.23}$$

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

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

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:

$$L_e = Q_e(I - M_e) \tag{3.24}$$

Where $L_e^*$ is eigenvalue matrix of $D^2L(\Theta^*)$. From equation 3.24, 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 3.25 specifies a so-called speed matrix $S_e^*$:

$$S_e^* = \begin{pmatrix} s_1^* = 1 - m_1^* & 0 & \cdots & 0 \\ 0 & s_2^* = 1 - m_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_r^* = 1 - m_r^* \end{pmatrix} \tag{3.25}$$

From equations 3.22 and 3.25, we have $0 \le s_i^* \le 1$. Equation 3.26 specifies $L_e^*$ which is eigenvalue matrix of $D^2L(\Theta^*)$.

$$L_e^* = \begin{pmatrix} l_1^* = q_1^* s_1^* & 0 & \cdots & 0 \\ 0 & l_2^* = q_2^* s_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_r^* = q_r^* s_r^* \end{pmatrix} \tag{3.26}$$

From equation 3.25, suppose $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_r^{(t)})$ at current $t^{\text{th}}$ 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 $\theta_i^*$. From equations 3.23 and 3.25, equation 3.27 specifies the speed $s^*$ of GEM algorithm.

$$s^* = 1 - m^* \tag{3.27}$$

Where,

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

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

For example, when $\Theta$ degrades into scalar as $\Theta = \theta$, 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 3.20 with $\theta^* = 0.6268$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

In practice, if GEM is run step by step, $\theta^*$ is not known yet at some $t^{\text{th}}$ iteration when GEM does not converge yet. Hence, equation 3.21 (McLachlan & Krishnan, 1997, p. 120) is used to make approximation of $M_e^* = DM(\theta^*)$ with unknown $\theta^*$ 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 $t^{\text{th}}$ 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 $t = 4$.

We will prove by contradiction that if definition 1 is satisfied strictly such that $Q(M(\Theta^{(t)}) \mid \Theta^{(t)}) > Q(\Theta^{(t)} \mid \Theta^{(t)})$ then, $l_i^* < 0$ for all $i$. Conversely, suppose we *always* have $l_i^* \geq 0$ for some $i$ when definition 1 is satisfied strictly. When $\Theta$ degrades into scalar as $\Theta = \theta$, we have converse assumption "$D^2L(\theta^*) = L_e^* \geq 0$" when definition 1 is satisfied strictly. In fact, when $Q(M(\Theta^{(t)}) \mid \Theta^{(t)}) > Q(\Theta^{(t)} \mid \Theta^{(t)})$, the sequence $\{L(\theta^{(t)})\}_{t=1}^{+\infty} = L(\theta^{(1)}), L(\theta^{(2)}), \ldots, L(\theta^{(t)}), \ldots$ is strictly increasing, which in turn causes that the sequence $\{\theta^{(t)}\}_{t=1}^{+\infty} = \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}, \ldots$ is strictly monotonous. This means:

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

Or

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

It implies

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

So we have

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

From equation 3.26, we deduce that $D^2L(\theta^*) = L_e^* = Q_e^* S_e^* < 0$ where $Q_e^* = D^{20}Q(\theta^* \mid \theta^*) < 0$. However, this contradict the converse assumption "$l_i^* \geq 0$ for some $i$ when definition 1 is satisfied strictly". Therefore, if definition 1 is satisfied strictly then, $l_i^* < 0$ for all $i$. In other words, at that time, $D^2L(\Theta^*) = L_e^*$ is negative definite. As a result, following is corollary 4.

**Corollary 4.** If an algorithm satisfies three following assumptions:

(1) $Q(M(\Theta^{(t)}) \mid \Theta^{(t)}) > Q(\Theta^{(t)} \mid \Theta^{(t)})$ for all $t$.

(2) $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite for all $t$.

(3) The sequence $\{L(\Theta^{(t)})\}_{t=1}^{+\infty}$ is bounded above.

Then,

(1) Such algorithm is an GEM and converges to a stationary point $\Theta^*$ of the likelihood function $L(\Theta)$ such that $DL(\Theta^*) = \mathbf{0}^T$ and $D^2L(\Theta^*)$ negative definite.

(2) $D^{20}Q(\Theta^* \mid \Theta^*)$ is negative definite.

(3) Equation 3.11 is retrieved ■

Note that $\Theta_0^{(t+1)}$ is on the line segment joining $\Theta^{(t)}$ and $\Theta^{(t+1)}$ according to equation 3.15.

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

Both $-D^{20}H(\Theta^*|\Theta^*)$ and $-D^{20}Q(\Theta^*|\Theta^*)$ are Fisher information matrices (Zivot, 2009, pp. 7-9) specified by equation 3.28.

$$I_H(\Theta^*) = -D^{20}H(\Theta^*|\Theta^*)$$
$$I_Q(\Theta^*) = -D^{20}Q(\Theta^*|\Theta^*) \quad (3.28)$$

$I_H(\Theta^*)$ measures information of $X$ about $\Theta^*$ with support of $Y$ whereas $I_Q(\Theta^*)$ measures information of $X$ about $\Theta^*$. 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 $\Theta^*$ 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 3.29.

$$V_H(\Theta^*) = \left(I_H(\Theta^*)\right)^{-1}$$
$$V_Q(\Theta^*) = \left(I_Q(\Theta^*)\right)^{-1} \quad (3.29)$$

Equation 3.30 is a variant of equation 3.11 to calculate $DM(\Theta^*)$ based on information matrices:

$$DM(\Theta^*) = I_H(\Theta^*)\left(I_Q(\Theta^*)\right)^{-1} = \left(V_H(\Theta^*)\right)^{-1}V_Q(\Theta^*) \quad (3.30)$$

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

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

And

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

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

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

Similarly, we have:

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

Hence, equation 3.31 specifies $DM(\Theta^*)$ in case of exponential family.

$$DM(\Theta^*) = V(\tau(X)|Y,\Theta^*)V(\tau(X)|\Theta^*) \quad (3.31)$$

## 4. Variants of EM algorithm

The main purpose of EM algorithm (GEM algorithm) is to maximize the log-likelihood $L(\Theta) = \log(g(Y|\Theta))$ with observed data (incomplete data) $Y$ by maximizing the condition expectation $Q(\Theta'|\Theta)$. Such $Q(\Theta'|\Theta)$ is defined fixedly in E-step. Therefore, most variants of EM algorithm focus on how to maximize $Q(\Theta'|\Theta)$ 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 $f(X|\Theta)$, $g(Y|\Theta)$ or $k(X|Y,\Theta)$ according to maximum a posteriori probability (MAP) method (Wikipedia, Maximum a posteriori estimation, 2017). For example, if $\pi(\Theta)$ in conjugation with $g(Y|\Theta)$ then, the posterior probability $\pi(\Theta|Y)$ is:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

Because $\int_{\Theta} g(Y|\Theta)\pi(\Theta)\mathrm{d}\Theta$ is constant with regard to $\Theta$, the optimal likelihood-maximization estimate $\Theta^*$ 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 3.24 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 4.1.2 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,

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

Because $Q(\Theta' \mid \Theta)$ and $\pi(\Theta')$ are smooth enough, $D^{20}Q(\Theta^* \mid \Theta^*)$ 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 $\Pi_e^*$ 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 3.17. Because $\pi(\Theta^*)$ is non-convex function, eigenvalues $\pi_1^*, \pi_2^*, \ldots, \pi_r^*$ of $\Pi_e^*$ 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:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

According to equation 3.16, 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 3.18 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.

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

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

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

Because all $h_i^*$, $q_i^*$, and $\pi_i^*$ 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 3.23, 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 $\Theta$. 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 $\Theta_0$ as a solution candidate and also goes through many iterations. Suppose the current parameter is $\Theta_i$, the next value $\Theta_{i+1}$ 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 $i^{\text{th}}$ iteration. At that time, $\Theta_{i+1}$ 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:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

Because $Q(\Theta \mid \Theta^{(t)})$ is smooth enough, $D^{20}Q(\Theta \mid \Theta^{(t)})$ is symmetric matrix according to Schwarz's theorem (Wikipedia, Symmetry of second derivatives, 2018), which imlies:

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

So we have:

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

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

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

It implies:

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

This means:

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

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 the condition is that $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ is negative definite and $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite.

*E-step*:

The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$, according to equation 2.4.

*M-step*:

The next parameter $\Theta^{(t+1)}$ is:

$$\Theta^{(t+1)} = \Theta^{(t)} - \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1}\left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T \quad (4.2.2)$$

**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. From equation 3.15, we have:

$$\begin{aligned} &Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) \\ &\quad = -D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta^{(t+1)} - \Theta^{(t)}\right) \\ &\quad - \left(\Theta^{(t+1)} - \Theta^{(t)}\right)^T D^{20}\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta^{(t+1)} - \Theta^{(t)}\right) \end{aligned}$$

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

$$\begin{aligned} &Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) \\ &= -D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) * \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1} * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T \\ &-D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) * \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1} * D^{20}\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right) * \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1} \\ &\quad * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T \end{aligned}$$

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

Let,

$$A = \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1} * D^{20}\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right) * \left(D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{-1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

Because $Q(\Theta' \mid \Theta)$ is smooth enough, $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ and $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ are symmetric matrices according to Schwarz's theorem (Wikipedia, Symmetry of second derivatives, 2018). Thus, $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ and $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ are commutative:

$$D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)}) = D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$$

Suppose both $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ and $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ 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):

$$D^{20}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) = WQ_e^{(t)}W^{-1}$$

$$D^{20}Q\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right) = WQ_e^{(t+1)}W^{-1}$$

Where $Q_e^{(t)}$ and $Q_e^{(t+1)}$ are eigenvalue matrices of $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ and $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$, respectively. Matrix $A$ is decomposed as below:

$$A = \left(WQ_e^{(t)}W^{-1}\right)^{-1} * \left(WQ_e^{(t+1)}W^{-1}\right) * \left(WQ_e^{(t)}W^{-1}\right)^{-1}$$

$$= W\left(Q_e^{(t)}\right)^{-1}W^{-1}WQ_e^{(t+1)}W^{-1}W\left(Q_e^{(t)}\right)^{-1} = W\left(Q_e^{(t)}\right)^{-1}Q_e^{(t+1)}Q_e^{(t)}W^{-1}$$

$$= W\left(Q_e^{(t)}\right)^{-1}Q_e^{(t)}Q_e^{(t+1)}W^{-1} = WQ_e^{(t+1)}W^{-1}$$

(Because $Q_e^{(t)}$ and $Q_e^{(t+1)}$ are commutative)

Hence, eigenvalue matrix of $A$ is also $Q_e^{(t+1)}$. Because $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite, $A$ is negative definite too. We have:

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

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

$$-D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) * A * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{T}$$

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

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

Because $A$ is negative definite, we have:

$$D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) * A * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^{T} < 0$$

As a result, we have:

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

Hence, EM1 surely converges to a local maximizer $\Theta^*$ according to corollary 4 on condition that $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ is negative definite and $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite with assumption that the sequence $\left\{L\left(\Theta^{(t)}\right)\right\}_{t=1}^{+\infty}$ is bounded above.

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.

$$M_e = H_eQ_e^{-1}$$

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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 a local maximizer $\Theta^*$ 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 strictly definition 1 of GEM algorithm. So the convergence to a local maximizer $\Theta^*$ of equation 4.2.4 is also asserted with condition that $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ is negative definite and $D^{20}Q(\Theta_0{}^{(t+1)} \mid \Theta^{(t)})$ is negative definite with assumption that the sequence $\left\{L\left(\Theta^{(t)}\right)\right\}_{t=1}^{+\infty}$ is bounded above. 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 $\omega(\theta)$. According to Ta (Ta, 2014), the first-order derivative of $\eta(\theta)$ is:

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

According to Ta (Ta, 2014), the second-order derivative of $\eta(\theta)$ is:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

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

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

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

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

In order to achieve $\eta''(\bar{\theta}) = 0$, Ta (Ta, 2014) selected:

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

According to Ta (Ta, 2014), Newton-Raphson process is improved as follows:

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

This means:

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

As a result, equation 4.2.4 is a generality of the equation above when $\Theta$ is vector.

I propose to apply gradient descent method (Ta, 2014) into M-step of GEM so that Newton-Raphson process is replaced by gradient descent process with expectation that descending direction which is the opposite of gradient vector $D^{10}Q(\Theta \mid \Theta^{(t)})$ speeds up convergence of GEM. Table 4.2.2 specifies GEM associated with gradient descent method, which is called GD-GEM algorithm. The condition is that $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite.

*E-step*:

The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$, according to equation 2.4.

*M-step*:

The next parameter $\Theta^{(t+1)}$ is:

$$\Theta^{(t+1)} = \Theta^{(t)} - \gamma^{(t)}\Big(D^{10}Q\big(\Theta^{(t)}\big|\Theta^{(t)}\big)\Big)^T \quad (4.2.5)$$

Where $\gamma^{(t)} > 0$ is length of the descending direction. As usual, $\gamma^{(t)}$ is selected such that

$$\gamma^{(t)} = \underset{\gamma}{\operatorname{argmax}}\, Q\big(\Phi^{(t)}\big|\Theta^{(t)}\big) \quad (4.2.6)$$

Where,

$$\Phi^{(t)} = \Theta^{(t)} + \gamma D^{10}Q\big(\Theta^{(t)}\big|\Theta^{(t)}\big)$$

**Table 4.2.1.** E-step and M-step of GD-GEM algorithm

Note, gradient descent method is used to solve minimization problem but its use for solving maximization problem is the same. From equation 3.15, we have:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

$$Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) = -D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta^{(t+1)} - \Theta^{(t)}\right) - \left(\Theta^{(t+1)} - \Theta^{(t)}\right)^T D^{20}\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta^{(t+1)} - \Theta^{(t)}\right)$$

By substituting equation 4.2.5 for $Q(\Theta^{(t+1)} \mid \Theta^{(t)}) - Q(\Theta^{(t+1)} \mid \Theta^{(t)})$, we have:

$$Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) = \gamma^{(t)} D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T - \left(\gamma^{(t)}\right)^2 D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) * D^{20}\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right) * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T$$

Due to:

$$D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) * \left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right)\right)^T \geq 0$$

$$D^{20}\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right) \text{ negative definite}$$

$$\gamma^{(t)} > 0$$

As a result, we have:

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

Hence, GD-GEM surely converges to a local maximizer $\Theta^*$ according to corollary 4 on condition that $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite with assumption that the sequence $\left\{L\left(\Theta^{(t)}\right)\right\}_{t=1}^{+\infty}$ is bounded above.

It is not easy to solve the maximization problem with regard to $\gamma$ according to equation 4.2.6. So if $Q(\Theta \mid \Theta^{(t)})$ satisfies Wolfe conditions (Wikipedia, Wolfe conditions, 2017) and concavity and $D^{10}Q(\Theta \mid \Theta^{(t)})$ is Lipschitz continuous (Wikipedia, Lipschitz continuity, 2018) then, equation 4.2.6 is replaced by equation 4.2.7 (Wikipedia, Gradient descent, 2018).

$$\gamma^{(t)} = \frac{\left(D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) - D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t-1)}\right)\right)\left(\Theta^{(t)} - \Theta^{(t-1)}\right)}{\left|D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) - D^{10}Q\left(\Theta^{(t)}\middle|\Theta^{(t-1)}\right)\right|^2} \quad (4.2.7)$$

Where |.| denotes length or module of 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 $\Theta$ is scalar such that $\Theta = \theta$, the sequence $\left\{\theta^{(t)}\right\}_{t=1}^{+\infty} = \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}, \ldots$ is monotonous. From equation 3.20

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

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, $\theta^*$ is approximated by (Lambers, 2009, p. 1)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

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

We construct Aitken sequence $\left\{\hat{\theta}^{(t)}\right\}_{t=1}^{+\infty} = \hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots, \hat{\theta}^{(t)}, \dots$ 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 $\Delta$ 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 $\Theta$ is vector as $\Theta = (\theta_1, \theta_2, \dots, \theta_r)^T$, Aitken sequence $\left\{\widehat{\Theta}^{(t)}\right\}_{t=1}^{+\infty} = \widehat{\Theta}^{(1)}, \widehat{\Theta}^{(2)}, \dots, \widehat{\Theta}^{(t)}, \dots$ is defined by applying equation 4.3.1 into its components $\theta_i$ (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 $\left\{\widehat{\Theta}^{(t)}\right\}_{t=1}^{+\infty}$ approaches $\Theta^*$ faster than the sequence $\left\{\Theta^{(t)}\right\}_{t=1}^{+\infty} = \Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(t)}, \dots$ with note that the sequence $\left\{\Theta^{(t)}\right\}_{t=1}^{+\infty}$ 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 $\left\{\Theta^{(t)}\right\}_{t=1}^{+\infty}$ by Aitken sequence $\left\{\widehat{\Theta}^{(t)}\right\}_{t=1}^{+\infty}$ as seen in table 4.3.1.

*E-step*:

The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$, according to equation 2.4. Note that $t = 1, 2, 3, \dots$ and $\Theta^{(0)} = \Theta^{(1)}$.

*M-step*:

Let $\Theta^{(t+1)} = (\theta_1^{(t+1)}, \theta_2^{(t)}, \dots, \theta_r^{(t+1)})^T$ 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 $\widehat{\Theta}^{(t-1)} = \left(\hat{\theta}_1^{(t-1)}, \hat{\theta}_2^{(t-1)}, \dots, \hat{\theta}_r^{(t-1)}\right)^T$ 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 $\left\{\widehat{\Theta}^{(t)}\right\}_{t=1}^{+\infty}$ converges to $\Theta^*$ faster than the sequence $\left\{\Theta^{(t)}\right\}_{t=1}^{+\infty}$ does, the convergence of GEM is improved with support of Aitken acceleration method.

In equation 4.3.2, parametric components $\theta_i$ (s) converges separately. Guo, Li, and Xu (Guo, Li, & Xu, 2017) assumed such components converges together with the same rate. So they

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

replaced equation 4.3.2 by equation 4.3.3 (Guo, Li, & Xu, 2017, p. 176) for Aitken sequence $\{\widehat{\Theta}^{(t)}\}_{t=1}^{+\infty}$.

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

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

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

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 $\Theta$ but with some function $g_s(\Theta)$ fixed at its previous value. Concretely, there are $S$ CM-steps and every $s^{\text{th}}$ CM-step finds $\Theta^{(t+s/S)}$ that maximizes $Q(\Theta \mid \Theta^{(t)})$ over $\Theta$ 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, $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$ according to equation 2.4.

*CM-steps*:

There are $S$ CM-steps. In every $s^{\text{th}}$ 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 ($S^{\text{th}}$ 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)$^{\text{th}}$ iteration).

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

ECM algorithm stops at some $t^{\text{th}}$ 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 $\Theta_s$ then, the $s^{\text{th}}$ CM-step maximizes $Q(\Theta \mid \Theta^{(t)})$ with regard to $\Theta_s$ whereas other sub-parameters are fixed. If $g_s(\Theta)$ consists only $\Theta_s$ then, the $s^{\text{th}}$ CM-step maximizes $Q(\Theta \mid \Theta^{(t)})$ with regard to all sub-parameters except $\Theta_s$. 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, the convergence of ECM is asserted according to corollary 3 on condition that $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite with assumption that the sequence $\{L(\Theta^{(t)})\}_{t=1}^{+\infty}$ is bounded above. However, Meng and Rubin (Meng & Rubin, 1993, pp. 274-276) provided some conditions for convergence of ECM to a maximizer of $L(\Theta)$.

## 5. Discussions

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

$\log(g(Y \mid \Theta))$ 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 $t^{\text{th}}$ iteration, which is also a maximizer of the lower bound at $t^{\text{th}}$ iteration. The lower bound is increased after every iteration. As a result, the maximizer $\Theta^{*}$ 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 $t^{\text{th}}$ 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 seen in table 5.1.

| |
|---|
| *E-step*:<br>The lower bound $lb(\Theta \mid \Theta^{(t)})$ is re-calculated based on $Q(\Theta \mid \Theta^{(t)})$.<br>*M-step*:<br>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. |

**Table 5.1.** An interpretation of GEM with lower bound

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.

The convergence of GEM is based on the assumption that $Q(\Theta' \mid \Theta)$ is smooth enough but $Q(\Theta' \mid \Theta)$ may not be smooth in practice when $f(X \mid \Theta)$ may be discrete probability function. For example, when $f(X \mid \Theta)$ and $k(X \mid Y, \Theta)$ are discrete, equation 2.4 becomes

$$Q(\Theta'|\Theta) = E\left(\log\left(f(X|\Theta')\right)\middle|Y,\Theta\right) = \sum_{X(Y)} k(X|Y,\Theta)\log\left(f(X|\Theta')\right)$$

This discussion section goes beyond traditional variants of GEM algorithm when $Q(\Theta' \mid \Theta)$ is not smooth. Therefore, heuristic optimization methods which simulate social behavior, such as particle swarm optimization (PSO) algorithm (Poli, Kennedy, & Blackwell, 2007) and artificial bee colony (ABC) algorithm, are useful in case that there is no requirement of existence of derivative. Moreover, these heuristic methods aim to find global optimizer. I propose an association of GEM and PSO which produces a so-called quasi-PSO-GEM algorithm in which M-step is implemented by one-time PSO (Wikipedia, Particle swarm optimization, 2017). Given current $t^{\text{th}}$ iteration, $\Theta^{(t)}$ is modeled as swarm's best position. Suppose there are $n$ particles and each particle $i$ has current velocity $V_i^{(t)}$, current positions $\Psi_i^{(t)}$, and best position $\Phi_i^{(t)}$. At each iteration, it is expected that these particles move to swarm's new best position which is the next parameter $\Theta^{(t+1)}$. The swarm's best position at the final iteration is expected as $\Theta^{*}$. Table 5.2 is the proposal of quasi-PSO-GEM algorithm.

| |
|---|
| *E-step*:<br>As usual, $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$ according to equation 2.4.<br>*M-step* includes four sub-steps: |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2018 doi:10.20944/preprints201802.0131.v2

1. Calculating the next velocity $V_i^{(t+1)}$ of each particle based on its current velocity $V_i^{(t)}$, its current positions $\Psi_i^{(t)}$, its best positions $\Phi_i^{(t)}$, and the swarm's best position $\Theta^{(t)}$:
$$V_i^{(t+1)} = \omega V_i^{(t)} + r\phi_1\left(\Phi_i^{(t)} - \Psi_i^{(t)}\right) + r\phi_2\left(\Theta^{(t)} - \Psi_i^{(t)}\right) \quad (5.1)$$
Where $\omega$, $\phi_1$, and $\phi_2$ are particular parameters of PSO (Poli, Kennedy, & Blackwell, 2007, pp. 3-4) whereas $r$ is a random number such that $0 < r < 1$ (Wikipedia, Particle swarm optimization, 2017).
2. Calculating the next position $\Psi_i^{(t+1)}$ of each particle based on its current position $\Psi_i^{(t)}$ and its current velocity $V_i^{(t)}$:
$$\Psi_i^{(t+1)} = \Psi_i^{(t)} + V_i^{(t)} \quad (5.2)$$
3. If $Q(\Phi_i^{(t)} \mid \Theta^{(t)}) < Q(\Psi_i^{(t+1)} \mid \Theta^{(t)})$ then, the next best position of each particle $i$ is re-assigned as $\Phi_i^{(t+1)} = \Psi_i^{(t+1)}$. Otherwise, such next best position is kept intact as $\Phi_i^{(t+1)} = \Phi_i^{(t)}$.
4. The next parameter $\Theta^{(t+1)}$ is the swarm's new best position over the best positions of all particles:
$$\Theta^{(t+1)} = \underset{\Phi_i^{(t)}}{\operatorname{argmax}} \left\{ Q\left(\Phi_1^{(t)} \middle| \Theta^{(t)}\right), Q\left(\Phi_2^{(t)} \middle| \Theta^{(t)}\right), \ldots, Q\left(\Phi_n^{(t)} \middle| \Theta^{(t)}\right) \right\} \quad (5.3)$$

If the bias $|\Theta^{(t+1)} - \Theta^{(t)}|$ is small enough, the algorithm stops. Otherwise, $\Theta^{(t+1)}$ and all $V_i^{(t+1)}$, $\Psi_i^{(t+1)}$, $\Phi_i^{(t+1)}$ become current parameters in the next iteration.

**Table 5.1.** E-step and M-step of the proposed quasi-PSO-GEM

At the first iteration, each particle is initialized with $\Psi_i^{(1)} = \Phi_i^{(1)} = \Theta^{(1)}$ and uniformly distributed velocity $V_i^{(1)}$. Note, $\Theta^{(1)}$ is initialized arbitrarily. Other termination criteria can be used, for example, $Q(\Theta \mid \Theta^{(t)})$ is large enough or the number of iterations is large enough.

We cannot prove mathematically convergence of quasi-PSO-GEM but we expect that $\Theta^{(t+1)}$ resulted from equation 5.3 is an approximation of $\Theta^*$ at the last iteration after a large enough number of iterations. However, quasi-PSO-GEM tendentiously approaches global maximizer of $L(\Theta)$, regardless of whether $L(\Theta)$ is concave. Hence, it is necessary to make experiment on quasi-PSO-GEM.

There are many other researches which combine EM and PSO but the proposed quasi-PSO-GEM algorithm has different ideology when it one-time PSO is embed into M-step to maximize $Q(\Theta \mid \Theta^{(t)})$ and so the ideology of quasi-PSO-GEM is near to the ideology Newton-Raphson process. With different viewpoint, some other researches combine EM and PSO in order to solving better a particular problem instead of improving EM itself. For example, Ari and Aksoy (Ari & Aksoy, 2010) used PSO to solve optimization problem of the clustering algorithm based on mixture model and EM. Rajeswari and Gunasundari (Rajeswari & Gunasundari, 2016) proposed EM for PSO based weighted clustering. Zhang, Zhuang, Gao, Luo, Ran, and Du (Zhang, et al., 2014) proposed a so-called PSO-EM algorithm to make optimum use of PSO in partial E-step in order solve the difficulty of integrals in normal compositional model. Golubovic, Olcan, and Kolundzija (Golubovic, Olcan, & Kolundzija, 2007) proposed a few modifications of the PSO algorithm which are applied to EM optimization of a broadside antenna array. Tang, Song, and Liu (Tang, Song, & Liu, 2014) proposed a hybrid clustering method based on improved PSO and EM clustering algorithm to overcome drawbacks of EM clustering algorithm. Tran, Vo, and Lee (Tran, Vo, & Lee, 2013) proposed a novel clustering algorithm for image segmentation by employing the arbitrary covariance matrices that uses PSO for the estimation of Gaussian mixture models.

## References

Ari, C., & Aksoy, S. (2010). Maximum Likelihood Estimation of Gaussian Mixture Models Using Particle Swarm Optimization. *The 20th International Conference on Pattern*

*Recognition (ICPR 2010)* (pp. 746-749). Istanbul: IEEE. Retrieved February 21, 2018, from www.cs.bilkent.edu.tr/~saksoy/papers/icpr10_clustering.pdf

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

Golubovic, R. M., Olcan, D. I., & Kolundzija, B. M. (2007). Particle Swarm Optimization Algorithm and Its Modifications Applied to EM Problems. In B. D. Milovanović (Ed.), *The 8th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services (ELSIKS 2007)* (pp. 427-430). Nis, Serbia: IEEE. doi:10.1109/TELSKS.2007.4376029

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

Poli, R., Kennedy, J., & Blackwell, T. (2007, June). Particle swarm optimization. (M. Dorigo, Ed.) *Swarm Intelligence, 1*(1), 33-57. doi:10.1007/s11721-007-0002-0

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

Rajeswari, J., & Gunasundari, R. (2016, December). EMPWC: Expectation Maximization with Particle Swarm Optimization based Weighted Clustering for Outlier Detection in Large Scale Data. (C.-H. Lien, & T.-L. Liao, Eds.) *International Journal of Control Theory and Applications (IJCTA), 9*(36), 517-531. Retrieved February 21, 2018, from http://serialsjournals.com/articlesview.php?volumesno_id=1131&article_id=14367&volumes_id=848&journals_id=268

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

Tang, Z., Song, Y.-Q., & Liu, Z. (2014). Medical Image Clustering Based on Improved Particle Swarm Optimization and Expectation Maximization Algorithm. *The 6th Chinese Conference on Pattern Recognition (CCPR 2014). II*, pp. 360-371. Changsha, China: Springer. doi:10.1007/978-3-662-45643-9_38

Tran, A.-K., Vo, Q.-N., & Lee, G. (2013). Maximum Likelihood Estimation of Gaussian Mixture Models Using PSO for Image Segmentation. In J. Chen, A. Cuzzocrea, & L. T. Yang (Ed.), *The 2013 IEEE 16th International Conference on Computational Science and Engineering (CSE 2013)* (pp. 501-507). Sydney, NSW, Australia: IEEE. doi:10.1109/CSE.2013.81

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. (2017, March 7). *Particle swarm optimization*. (Wikimedia Foundation) Retrieved April 8, 2017, from Wikipedia website: https://en.wikipedia.org/wiki/Particle_swarm_optimization

Wikipedia. (2017, May 8). *Wolfe conditions*. (Wikimedia Foundation) Retrieved February 20, 2018, from Wikipedia website: https://en.wikipedia.org/wiki/Wolfe_conditions

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 28). *Gradient descent*. (Wikimedia Foundation) Retrieved February 20, 2018, from Wikipedia website: https://en.wikipedia.org/wiki/Gradient_descent

Wikipedia. (2018, February 17). *Lipschitz continuity*. (Wikimedia Foundation) Retrieved February 20, 2018, from Wikipedia website: https://en.wikipedia.org/wiki/Lipschitz_continuity

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

 doi:10.20944/preprints201802.0131.v2

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

Zhang, B., Zhuang, L., Gao, L., Luo, W., Ran, Q., & Du, Q. (2014, May 14). PSO-EM: A Hyperspectral Unmixing Algorithm Based On Normal Compositional Model. (A. Plaza, Ed.) *IEEE Transactions on Geoscience and Remote Sensing, 52*(12), 7782 - 7792. doi:10.1109/TGRS.2014.2319337

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.