Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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, variables are often denoted by letters such as x, y, z, X, Y, and Z whereas values and constants are often denoted by letters such as a, b, c, A, B, and C. Parameters are often denoted as Greek letters such as α, β, γ, Θ, Φ, and Ψ. Uppercase letters often denote vectors and matrices (multivariate quantities) whereas lowercase letters often denote scalars (univariate quantities). Script letters such as $\mathcal{X}$ and $\mathcal{Y}$ often denote data samples. Bold and uppercase letters such as $\boldsymbol{X}$ and $\boldsymbol{R}$ often denote algebraic structures such as spaces, fields, and domains. Moreover, bold and lowercase letters such as $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$, $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ may denote vectors. Bold and uppercase letters such as $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$, $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ may denote matrices.

By default, vectors are column vectors although a vector can be column vector or row vector. For example, given two vectors X and Y and two matrices A and B:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

X and *Y* above are column vectors. A row vector is represented as follows:

$$Z = (z_1, z_2, \ldots, z_r)$$

The number of elements in vector is its dimension. Zero vector is denoted as **0** whose dimension depends on context.

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

If considering rows and columns, $m_x n$ matrix *A* can be denoted A_{mxn} or $(a_{ij})_{mxn}$. Vector is 1-row matrix or 1-column matrix such as A_{1xn} or A_{nx1}. Scalar is 1-element vector or 1_x1 matrix. A matrix can be considered as a vector whose elements are vectors.

Let (**0**) denote zero matrix whose numbers of rows and columns depend on context. If considering rows and columns, zero matrix can be denoted $(0)_{mxn}$.

$$(\mathbf{0}) = (0)_{mxn} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

Matrix *A* is square if $m = n$, which can be denoted A_n or $(a_{ij})_n$. Matrix Λ is diagonal if it is square and its elements outside the main diagonal are zero:

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

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

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

Note, *I* is diagonal and its diagonal elements are 1. The row (column) number of identity matrix depends on context, but it can be denoted explicitly as I_n.

Vector addition and matrix addition are defined like numerical addition:

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

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

(if $n = k$)

Vector and matrix can be multiplied with a scalar.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$kX = \begin{pmatrix} kx_1 \\ kx_2 \\ \vdots \\ kx_r \end{pmatrix}$$

$$kA = \begin{pmatrix} ka_{11} & ka_{12} & \cdots & ka_{1n} \\ ka_{21} & ka_{22} & \cdots & ka_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ ka_{m1} & ka_{m2} & \cdots & ka_{mn} \end{pmatrix}$$

Let superscript "^{T}" denote transposition operator for vector and matrix, as follows:

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

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

Transposition operator is linear with addition operator as follows:

$$(X + Y)^T = X^T + Y^T$$

$$(A + B)^T = A^T + B^T$$

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

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

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

$$XY^T = YX^T = \begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_r \\ x_2y_1 & x_2y_2 & \cdots & x_2y_r \\ \vdots & \vdots & \ddots & \vdots \\ x_ry_1 & x_ry_2 & \cdots & x_ry_r \end{pmatrix}$$

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 notation |.| also denotes absolute value of scalar and determinant of square matrix; for example, we have $|-1| = 1$ and $|A|$ which is determinant of given square matrix A. Note, determinant is only defined for square matrix. If A has nonzero determinant ($\neq 0$), there exists its inverse denoted A^{-1} such that:

$$AA^{-1} = A^{-1}A = I$$

Where I is identity matrix. If matrix A has its inverse, A is called invertible or non-singular. In general, square matrix A is invertible is equivalent to the event that its determinant is nonzero ($\neq 0$). There are many documents which guide to calculate inverse of invertible matrix.

Let A and B be two invertible matrices, we have:

$$(AB)^{-1} = B^{-1}A^{-1}$$

$$|A^{-1}| = |A|^{-1} = 1 / |A|$$

Given invertible matrix A, it is called orthogonal matrix if $A^{-1} = A^T$, which means $AA^{-1} = A^{-1}A = AA^T = A^TA = I$.

Product (multiplication operator) of two matrices A_{mxn} and B_{nxk} is:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

Square 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 with the same rows and column then, they are commutative such that $AB = BA$ with note that the product AB and BA produces a symmetric matrix. Given invertible matrix A, if it is symmetric, its inverse A^{-1} is symmetric too.

Given N matrices A_i such that their product (multiplication operator) is valid, we have:

$$\left(\prod_{i=1}^{N} A_i\right)^T = (A_1 A_2 \dots A_N)^T = \prod_{i=N}^{1} A_i^T = A_N^T A_{N-1}^T \dots A_1^T$$

Product of matrix and vector is similar to product of matrix and matrix when vector is considered as 1-column matrix or 1-row matrix, which results a vector.

$$AX = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix}$$

Where $c_i = \sum_{j=1}^{n} a_{ij} x_j$.

$$Z^T A = (z_1, z_2, \dots, z_m) \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = C = (c_1, c_2, \dots, c_n)$$

Where $c_j = \sum_{i=1}^{m} z_i a_{ij}$.

Given square matrix A, tr(A) is trace operator which takes sum of its diagonal elements.

$$\mathrm{tr}(A) = \sum_{i} a_{ii}$$

Given invertible matrix A (n rows and n columns), Jordan decomposition theorem (Hardle & Simar, 2013, p. 63) stated that A is always decomposed as follows:

$$A = U \Lambda U^{-1} = U \Lambda U^T$$

Where U is orthogonal matrix composed of eigenvectors. Hence, U is called eigenvector matrix.

$$U = \begin{pmatrix} u_{11} & u_{21} & \dots & u_{n1} \\ u_{12} & u_{22} & \dots & u_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ u_{1n} & u_{2n} & \dots & u_{nn} \end{pmatrix}$$

There are n column eigenvectors $\boldsymbol{u}_i = (u_{11}, u_{12}, \dots, u_{1n})$ in U and they are mutually orthogonal, $\boldsymbol{u}_i^T \boldsymbol{u}_j = 0$. Where Λ is diagonal matrix composed of eigenvalues. Hence, Λ is called eigenvalue matrix.

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

Where λ_i are eigenvalues. When invertible matrix A is decomposed according to Jordan decomposition, we call A is diagonalized. If A can be diagonalized, it is called diagonalizable matrix. Of course, if A is invertible, A is diagonalizable. There are many documents for matrix diagonalization.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

Given two diagonalizable matrices A and B are equal size ($n_x n$) then, they are simultaneously diagonalizable (Wikipedia, Commuting matrices, 2017) and hence, there exists an orthogonal eigenvector matrix U such that (Wikipedia, Diagonalizable matrix, 2017) (StackExchange, 2013):

$$A = U\Gamma U^{-1} = U\Gamma U^T$$
$$B = U\Lambda U^{-1} = U\Lambda U^T$$

Where Γ and Λ are eigenvalue matrices of A and B, respectively.

Given symmetric matrix A, it is positive (negative) definite if and only if $X^TAX > 0$ ($X^TAX < 0$) for all vector $X \neq \mathbf{0}^T$. It is positive (negative) semi-definite if and only if $X^TAX \geq 0$ ($X^TAX \leq 0$) for all vector X. When diagonalizable A is diagonalized into $U\Lambda U^T$, it is positive (negative) definite if and only if all eigenvalues in Λ are positive (negative). Similarly, it is positive (negative) semi-definite if and only if all eigenvalues in Λ are non-negative (non-positive). If A is degraded as a scalar, concepts "positive definite", "positive semi-definite", "negative definite", and "negative semi-definite" becomes concepts "positive", "non-negative", "negative", and "non-positive", respectively.

Suppose $f(X)$ is scalar-by-vector function, for instance, 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 $\frac{\partial f(X)}{\partial x_i}$ is partial first-order derivative of f with regard to x_i. So gradient is row vector. The second-order derivative of $f(X)$ is called Hessian matrix as follows:

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

Where

$$\frac{\partial^2 f(X)}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left(\frac{\partial f(X)}{\partial x_j}\right)$$
$$\frac{\partial^2 f(X)}{\partial x_i^2} = \frac{\partial^2 f(X)}{\partial x_i \partial x_i}$$

Hence, second-order partial derivatives of x_i (s) are on diagonal of the Hessian matrix.

Hessian matrix is square matrix. Function $f(X)$ is called n^{th}-order analytic function or n^{th}-order smooth function if there is existence and continuity of k^{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.

$$\frac{\partial^2 f(X)}{\partial x_i \partial x_j} = \frac{\partial^2 f(X)}{\partial x_j \partial x_i}$$

When X is univariate, gradient vector and Hessian matrix are degraded as scalar values. Without loss of generality, by default, variable X in this research is multivariate as vector if there is no additional explanation.

Given $f(X)$ being second-order smooth function, $f(X)$ is convex (strictly convex) in domain $\boldsymbol{X}$ if and only if its Hessian matrix is semi-positive definite (positive definite) in $\boldsymbol{X}$. Similarly, $f(X)$ is concave (strictly concave) in domain $\boldsymbol{X}$ if and only if its Hessian matrix is semi-negative

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

definite (negative definite) in $\boldsymbol{X}$. Extreme point, optimized point, optimal point, or optimizer X^* is minimum point (minimizer) of convex function and is maximum point (maximizer) of concave function.

$$X^* = \underset{X \in \boldsymbol{X}}{\operatorname{argmin}} f(X) \text{ if } f \text{ convex in } \boldsymbol{X}$$

$$X^* = \underset{X \in \boldsymbol{X}}{\operatorname{argmax}} f(X) \text{ if } f \text{ concave in } \boldsymbol{X}$$

Given second-order smooth function $f(X)$, function $f(X)$ has stationary point X^* if its gradient vector at X^* is zero, $Df(X^*) = \mathbf{0}^T$. The stationary point X^* is local minimum point if Hessian matrix at X^* that is $D^2f(X^*)$ is positive definite. Otherwise, the stationary point X^* is local maximum point if Hessian matrix at X^* that is $D^2f(X^*)$ is negative definite. If a stationary point X^* is neither minimum point nor maximum point, it is saddle point in which $Df(X^*) = \mathbf{0}^T$ and $D^2f(X^*) = (\mathbf{0})$ where $(\mathbf{0})$ denotes zero matrix whose all elements are zero. Finding extreme point (minimum point or maximum point) is optimization problem. Therefore, if $f(X)$ is second-order smooth function and its gradient vector $Df(X)$ and Hessian matrix $D^2f(X)$ and are both determined, the optimization problem is processed by solving the equation created from setting the gradient $Df(X)$ to be zero ($Df(X)=\mathbf{0}^T$) and then checking if the Hessian matrix $Df(X^*)$ is positive definite or negative definite where X^* is solution of equation $Df(X)=\mathbf{0}^T$. If such equation cannot be solved due to its complexity, there are some popular methods to solve optimization problem such as Newton-Raphson (Burden & Faires, 2011, pp. 67-71) and gradient descent (Ta, 2014).

A short description of Newton-Raphson method is necessary because it is helpful to solve the equation $Df(X)=\mathbf{0}^T$ for optimization problem in practice, especially in case that there is no algebraic formula for solution of such equation. Suppose $f(X)$ is second-order smooth function, according to first-order Taylor series expansion of $Df(X)$ at $X=X_0$ with very small residual, we have:

$$Df(X) \approx Df(X_0) + (X - X_0)^T \left(D^2 f(X_0)\right)^T$$

Because $f(X)$ is second-order smooth function, $D^2f(X_0)$ is symmetric matrix according to Schwarz's theorem (Wikipedia, Symmetry of second derivatives, 2018), which implies:

$$D^2f(X_0) = (D^2f(X_0))^T$$

So, we have:

$$Df(X) \approx Df(X_0) + (X - X_0)^T D^2 f(X_0)$$

We expect that $Df(X) = \mathbf{0}^T$ so that X is a solution.

$$\mathbf{0}^T = Df(X) \approx Df(X_0) + (X - X_0)^T D^2 f(X_0)$$

It implies:

$$X^T \approx X_0^T - Df(X_0)\left(D^2 f(X_0)\right)^{-1}$$

This means:

$$X \approx X_0 - \left(D^2 f(X_0)\right)^{-1} \left(Df(X_0)\right)^T$$

Therefore, Newton-Raphson method starts with an arbitrary value of X_0 as a solution candidate and then goes through some iterations. Suppose at k^{th} iteration, the current value is X_k and the next value X_{k+1} is calculated based on following equation:

$$X_{k+1} \approx X_k - \left(D^2 f(X_k)\right)^{-1} \left(Df(X_k)\right)^T$$

The value X_k is solution of $Df(X)=\mathbf{0}^T$ if $Df(X_k)=\mathbf{0}^T$ which means that $X_{k+1}=X_k$ after some iterations. At that time $X_{k+1} = X_k = X^*$ is the local optimized point (local extreme point). So, the terminated condition of Newton-Raphson method is $Df(X_k)=\mathbf{0}^T$. Note, the X^* resulted from Newton-Raphson method is local minimum point (local maximum point) if $f(X)$ is convex function (concave function) in current domain.

Newton-Raphson method computes second-order derivative $D^2f(X)$ but gradient descent method (Ta, 2014) does not. This difference is not significant but a short description of gradient

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

descent method is necessary because it is also an important method to solve the optimization problem in case that solving directly the equation $Df(X)=\mathbf{0}^T$ is too complicated. Gradient descent method is also iterative method starting with an arbitrary value of X_0 as a solution candidate. Suppose at k^{th} iteration, the next candidate point X_{k+1} is computed based on the current X_k as follows (Ta, 2014):

$$X_{k+1} = X_k + t_k \boldsymbol{d}_k$$

The direction $\boldsymbol{d}_k$ is called descending direction, which is the opposite of gradient of $f(X)$. Hence, we have $\boldsymbol{d}_k = -Df(X_k)$. The value t_k is the length of the descending direction $\boldsymbol{d}_k$. The value t_k is often selected an minimizer (maximizer) of function $g(t) = f(X_k + t\boldsymbol{d}_k)$ for minimization (maximization) where X_k and $\boldsymbol{d}_k$ are known at k^{th} iteration. Alternately, t_k is selected by some advanced condition such as Barzilai–Borwein condition (Wikipedia, Gradient descent, 2018). After some iterations, point X_k converges to the local optimizer X^* when $\boldsymbol{d}_k = \mathbf{0}^T$. At that time is we have $X_{k+1} = X_k = X^*$. So, the terminated condition of Newton-Raphson method is $\boldsymbol{d}_k=\mathbf{0}^T$. Note, the X^* resulted from gradient descent method is local minimum point (local maximum point) if $f(X)$ is convex function (concave function) in current domain.

In the case that the optimization problem has some constraints, Lagrange duality (Jia, 2013) is applied to solve this problem. Given first-order smooth function $f(X)$ and constraints $g_i(X) \leq 0$ and $h_j(X) = 0$, the optimization problem is stated as follows:

$$\text{Optimize } f(X)$$
$$g_i(X) \leq 0 \text{ for } i = \overline{1,m}$$
$$h_j(X) = 0 \text{ for } j = \overline{1,n}$$

A so-called Lagrange function $la(X, \lambda, \mu)$ is established as sum of $f(X)$ and constraints multiplied by Lagrange multipliers λ and μ. In case of minimization problem, $la(X, \lambda, \mu)$ is

$$la(X,\lambda,\mu) = f(X) + \sum_{i=1}^{m} \lambda_i g(X) + \sum_{j=1}^{n} \mu_j h(X)$$

In case of maximization problem, $la(X, \lambda, \mu)$ is

$$la(X,\lambda,\mu) = f(X) - \sum_{i=1}^{m} \lambda_i g(X) - \sum_{j=1}^{n} \mu_j h(X)$$

Where all $\lambda_i \geq 0$. Note, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)^T$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_m)^T$ are called Lagrange multipliers and $la(X, \lambda, \mu)$ is function of X, λ, and μ. Thus, optimizing $f(X)$ with subject to constraints $g_i(X) \leq 0$ and $h_j(X) = 0$ is equivalent to optimize $la(X, \lambda, \mu)$, which is the reason that this method is called Lagrange duality. Suppose $la(X, \lambda, \mu)$ is also first-order smooth function. In case of minimization problem, the gradient of $la(X, \lambda, \mu)$ with regard to X is

$$Dla(X,\lambda,\mu) = Df(X) + \sum_{i=1}^{m} \lambda_i Dg(X) + \sum_{j=1}^{n} \mu_j Dh(X)$$

In case of maximization problem, the gradient of $la(X, \lambda, \mu)$ with regard to X is

$$Dla(X,\lambda,\mu) = Df(X) - \sum_{i=1}^{m} \lambda_i Dg(X) - \sum_{j=1}^{n} \mu_j Dh(X)$$

According to KKT condition (Wikipedia, Karush–Kuhn–Tucker conditions, 2014), a local optimized point (local extreme point) X^* is solution of the following equation system:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\begin{cases} Dla(X, \lambda, \mu) = \mathbf{0}^T \\ g_i(X) \leq 0 \text{ for } i = \overline{1, m} \\ h_j(X) = 0 \text{ for } j = \overline{1, n} \\ \lambda_i \geq 0 \text{ for } i = \overline{1, m} \\ \sum_{i=1}^{m} \lambda_i g(X) = 0 \end{cases}$$

The last equation in the KKT system above is called complementary slackness. The main task of KKT problem is to solve the first equation $Dla(X, \lambda, \mu) = \mathbf{0}^T$. Again some practical methods such as Newton-Raphson method can be used to solve the equation $Dla(X, \lambda, \mu) = \mathbf{0}^T$. Alternately, gradient descent method can be used to optimize $la(X, \lambda, \mu)$ with constraints specified in KKT system.

$$\begin{cases} g_i(X) \leq 0 \text{ for } i = \overline{1, m} \\ h_j(X) = 0 \text{ for } j = \overline{1, n} \\ \lambda_i \geq 0 \text{ for } i = \overline{1, m} \\ \sum_{i=1}^{m} \lambda_i g(X) = 0 \end{cases}$$

Let $P(.)$ denote probability,

$$0 \leq P(.) \leq 1$$

We need to skim some essential probabilistic rules such as additional rule, multiplication rule, total probability rule, and Bayes' rule. Given two random events (or random variables) X and Y, additional rule (Montgomery & Runger, 2003, p. 33) and multiplication rule (Montgomery & Runger, 2003, p. 44) are expressed as follows:

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

$$P(X \cap Y) = P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

Where notations $\cup$ and $\cap$ denote union operator and intersection operator in set theory (Wikipedia, Set (mathematics), 2014). Your attention please, when X and Y are numerical variables, notations $\cup$ and $\cap$ also denote operators "or" and "and" in theory logic (Rosen, 2012, pp. 1-12). The probability $P(X, Y)$ is known as joint probability. The probability $P(X|Y)$ is called conditional probability of X given Y:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} = \frac{P(X \cap Y)}{P(Y)} = \frac{P(Y|X)P(X)}{P(Y)}$$

Conditional probability is base of Bayes' rule mentioned later.

If X and Y are mutually exclusive ($X \cap Y = \emptyset$) then, $X \cup Y$ is often denoted as $X+Y$ and we have:

$$P(X + Y) = P(X) + P(Y)$$

(Due to $P(\emptyset) = 0$)

X and Y are mutually independent if and only if one of three following conditions is satisfied:

$$P(X \cap Y) = P(X)P(Y)$$

$$P(X|Y) = P(X)$$

$$P(Y|X) = P(Y)$$

When X and Y are mutually independent, $X \cap Y$ are often denoted as XY and we have:

$$P(XY) = P(X, Y) = P(X \cap Y) = P(X)P(Y)$$

Given a complete set of mutually exclusive events X_1, X_2,…, X_n such that

$$X_1 \cup X_2 \cup \ldots \cup X_n = X_1 + X_2 + \cdots + X_n = \Omega \text{ where } \Omega \text{ is probability space}$$

$$X_i \cap X_j = \emptyset, \forall i, j$$

The total probability rule (Montgomery & Runger, 2003, p. 44) is specified as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$P(Y) = P(Y|X_1)P(X_1) + P(Y|X_2)P(X_2) + \cdots + P(Y|X_n)P(X_n) = \sum_{i=1}^{n} P(Y|X_i)P(X_i)$$

$$\text{Where } X_1 + X_2 + \cdots + X_n = \Omega \text{ and } X_i \cap X_j = \emptyset, \forall i, j$$

If X and Y are continuous variables, the total probability rule is re-written in integral form as follows:

$$P(Y) = \int_X P(Y|X)P(X)\mathrm{d}X$$

Note, $P(Y|X)$ and $P(X)$ are continuous functions known as probability density functions mentioned later. The important Bayes' rule will also be mentioned later.

A variable X is called random variable if it conforms a probabilistic distribution which is specified by a probability density function (PDF) or a cumulative distribution function (CDF) (Montgomery & Runger, 2003, p. 64) (Montgomery & Runger, 2003, p. 102) but CDF and PDF have the same meaning and they share interchangeable property when PDF is derivative of CDF; in other words, CDF is integral of PDF. In practical statistics, PDF is used more common than CDF is used and so, PDF is mentioned over the whole report. When X is discrete, PDF is degraded as probability of X. Note, notation $P(.)$ often denotes probability and it can be used to denote PDF but we prefer to use lower case letters such as f and g to denote PDF. Given a random variable having PDF f, we often state that "such variable has distribution f or such variable has density function f". Let $F(X)$ and $f(X)$ be CDF and PDF, respectively, equation 1.1 is definition of CDF and PDF.

$$\text{Continuous case: } \begin{cases} F(X_0) = P(X \le X_0) = \int_{-\infty}^{X_0} f(X)dX \\ \int_{-\infty}^{+\infty} f(X)\mathrm{d}X = 1 \end{cases}$$

$$\text{Discrete case: } \begin{cases} F(X_0) = P(X \le X_0) = \sum_{X \le X_0} P(X) \\ f(X) = P(X) \text{ and } \sum_X P(X) = 1 \end{cases} \quad (1.1)$$

In discrete case, probability at a single point X_0 is determined as $P(X_0) = f(X_0)$ but in continuous case, probability is determined in an interval $[a, b]$, (a, b), $[a, b)$, or $(a, b]$ where a, b, and X are real as integral of the PDF in such interval as follows:

$$P(a \le X \le b) = \int_a^b f(X)dX$$

Hence, in continuous case, probability at a single point is 0.

Equation 1.1 defines CDF and PDF for univariate random variable and so it is easy to expend it for multivariate variable when X is vector. Let $X = (x_1, x_2, \ldots, x_n)^T$ be n-dimension random vector, its CDF and PDF are re-defined as follows:

Continuous case: (1.2)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$F(X_0) = P(X \le X_0) = P(x_1 \le x_{01}, x_2 \le x_{02}, \dots, x_n \le x_{0n}) = \int_{-\infty}^{X_0} f(X)dX$$

$$= \int_{-\infty}^{X_0} \int_{-\infty}^{X_0} \dots \int_{-\infty}^{X_0} f(X)\mathrm{d}x_1\mathrm{d}x_2 \dots \mathrm{d}x_n$$

$$\int_{-\infty}^{+\infty} f(X)dX = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(X)\mathrm{d}x_1\mathrm{d}x_2 \dots \mathrm{d}x_n = 1$$

Discrete case:

$$F(X_0) = P(X \le X_0) = P(x_1 \le x_{01}, x_2 \le x_{02}, \dots, x_n \le x_{0n}) = \sum_{X \le X_0} P(X)$$

$$= \sum_{x_1 \le x_{01}} \sum_{x_2 \le x_{02}} \dots \sum_{x_n \le x_{0n}} P(X)$$

$$f(X) = P(X)$$

$$\sum_{X} P(X) = \sum_{x_1 \le x_{01}} \sum_{x_2 \le x_{02}} \dots \sum_{x_n \le x_{0n}} P(X) = 1$$

Marginal PDF of partial variable x_i where x_i is a component of X is the integral of $f(X)$ over all x_j except x_i.

$$f_{x_i}(x_i) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(X)\mathrm{d}x_1 \dots \mathrm{d}x_{i-1}\mathrm{d}x_{i+1} \dots \mathrm{d}x_n$$

Where,

$$\int_{-\infty}^{+\infty} f_{x_i}(x_i)\mathrm{d}x_i = 1$$

Joint PDF of x_i and x_j is defined as the integral of $f(X)$ over all x_k except x_i and x_j.

$$f_{x_i x_j}(x_i, x_j) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(X)\mathrm{d}x_1 \dots \mathrm{d}x_{i-1}\mathrm{d}x_{i+1} \dots \mathrm{d}x_{j-1}\mathrm{d}x_{j+1} \dots \mathrm{d}x_n$$

Where,

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{x_i x_j}(x_i, x_j)\,\mathrm{d}x_i\mathrm{d}x_j = 1$$

Conditional PDF of x_i given x_j is defined as follows:

$$f_{x_i|x_j}(x_i) = \frac{f_{x_i x_j}(x_i, x_j)}{f_{x_j}(x_j)}$$

Indeed, conditional PDF implies conditional probability.

Given random variable X and its PDF $f(X)$, theoretical expectation $E(X)$ and theoretical variance $V(X)$ of X are:

$$E(X) = \int_X Xf(X)\mathrm{d}X \tag{1.3}$$

$$V(X) = E\big(X - E(X)\big)\big(X - E(X)\big)^T = \int_X \big(X - E(X)\big)\big(X - E(X)\big)^T f(X)\mathrm{d}X$$
$$= E(XX^T) - E(X)E(X)^T \tag{1.4}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

Given two random variables X and Y along with a joint PDF $f(X, Y)$, theoretical covariance of X and Y is defined as follows:

$$V(X,Y) = E\big(X - E(X)\big)\big(Y - E(Y)\big)^T = \int_X \int_Y \big(X - E(X)\big)\big(Y - E(Y)\big)^T f(X,Y)\mathrm{d}X\mathrm{d}Y \quad (1.5)$$

If the random variables X and Y are mutually independent given the joint PDF $f(X, Y)$, its covariance is zero, $V(X, Y)=0$. Note, joint PDF is the PDF having two or more random variables. When X and Y are multivariate vectors, $V(X, Y)$ is covariance matrix of X and Y given the joint PDF $f(X, Y)$.

The expectation $E(X)$ of X is often called theoretical mean. When X is multivariate vector, $E(X)$ is mean vector and $V(X)$ is covariance matrix. Note, covariance matrix is always symmetric and invertible. As usual, $E(X)$ and $V(X)$ are often denoted as μ and Σ, respectively if they are parameters of PDF. When X is univariate, $E(X)$ and $V(X)$ are scalars and $V(X)$ is often denoted σ^2 (if it is parameter of PDF). For example, if X is univariate and follows normal distribution, its PDF is:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}\right)$$

If X is multivariate and follows multivariate normal distribution, its PDF is:

$$f(X) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

When $X = (x_1, x_2, \ldots, x_n)^T$ is multivariate, μ and Σ have following forms:

$$\mu = \begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{pmatrix}$$

$$\Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}\end{pmatrix}$$

Of course, μ and Σ are determined by equation 1.3 and equation 1.4, respectively. However, theoretical means and variances of partial variables x_i can be determined separately. For instance, each μ_j is theoretical mean of partial variable x_j given marginal PDF $f_{x_j}(x_j)$.

$$\mu_i = E(x_i) = \int_{x_i} x_i f_{x_i}(x_i)\mathrm{d}x_i$$

Each σ_{ij} where $i \neq j$ is theoretical covariance of partial variables x_i and x_j given joint PDF $f_{x_i x_j}(x_i, x_j)$.

$$\sigma_{ij} = V(x_i, x_j) = E(x_i - \mu_i)(x_j - \mu_j) = \int_{x_i}\int_{x_j}(x_i - \mu_i)(x_j - \mu_j) f_{x_i x_j}(x_i, x_j)\mathrm{d}x_i\mathrm{d}x_j$$

Each σ_{ii} on diagonal of Σ is theoretical variance of partial variable x_i given marginal PDF $f_{x_j}(x_j)$.

$$\sigma_{ii} = \sigma_i^2 = V(x_i) = E(x_i - \mu_i)^2 = \int_{x_i}(x_i - \mu_i)^2 f_{x_i}(x_i)\mathrm{d}x_i$$

Without loss of generality, by default, random variable X in this research is multivariate as vector if there is no additional explanation. Followings are some formulas related to theoretical expectation $E(X)$ and variance $V(X)$.

Let a and A be scalar constant and vector constant, respectively, we have:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$E(aX + A) = aE(X) + A$$
$$V(aX + A) = a^2V(X)$$

Given a set of random variables $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ and N scalar constants c_i (s), we have:

$$E\left(\sum_{i=1}^{N} c_i X_i\right) = \sum_{i=1}^{N} c_i E(X_i)$$
$$V\left(\sum_{i=1}^{N} c_i X_i\right) = \sum_{i=1}^{N} c_i^2 V(X_i) + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_i c_j V(X_i, X_j)$$

Where $V(X_i, X_j)$ is covariance of X_i and X_j.

If all X_i (s) are mutually independent, then

$$E\left(\sum_{i=1}^{N} c_i X_i\right) = \sum_{i=1}^{N} c_i E(X_i)$$
$$V\left(\sum_{i=1}^{N} c_i X_i\right) = \sum_{i=1}^{N} c_i^2 V(X_i)$$

If all X_i (s) are identically distributed, which implies that all X_i (s) are represented by the same random variable X, then

$$E\left(\sum_{i=1}^{N} c_i X_i\right) = \left(\sum_{i=1}^{N} c_i\right) E(X)$$
$$V\left(\sum_{i=1}^{N} c_i X_i\right) = \left(\sum_{i=1}^{N} c_i^2\right) V(X) + 2\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_i c_j\right) V(X)$$

If all X_i (s) are mutually independent and identically distributed (iid), then

$$E\left(\sum_{i=1}^{N} c_i X_i\right) = \left(\sum_{i=1}^{N} c_i\right) E(X)$$
$$V\left(\sum_{i=1}^{N} c_i X_i\right) = \left(\sum_{i=1}^{N} c_i^2\right) V(X)$$

Because EM algorithm is essentially an advanced version of maximum likelihood estimation (MLE) method, it is necessary to describe MLE in short. Suppose random variable X conforms to a distribution specified by the PDF denoted $f(X \mid \Theta)$ with parameter Θ. For example, if X is vector and follows normal distribution then,

$$f(X|\Theta) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Where μ and Σ are theoretical mean vector and covariance matrix, respectively with note that $\Theta = (\mu, \Sigma)^T$. The notation $|.|$ denotes determinant of given matrix and the notation Σ^{-1} denotes inverse of matrix Σ. Note, Σ is invertible and symmetric. Parameter of normal distribution is theoretical mean and theoretical variance,

$$\mu = E(X)$$
$$\Sigma = V(X) = E(X-\mu)(X-\mu)^T$$

But parameters of different distributions may be different from such mean and variance. Anyhow theoretical mean and theoretical variance can be calculated based on parameter Θ.

For example, suppose $X = (x_1, x_2, \ldots, x_n)^T$ follows multinomial distribution of K trials, its PDF is:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$f(X|\Theta) = \frac{K!}{\prod_{j=1}^{n}(x_j!)}\prod_{k=1}^{n} p_j^{x_j}$$

Where x_j are integers and $\Theta = (p_1, p_2, \ldots, p_n)^T$ is the set of probabilities such that

$$\sum_{j=1}^{n} p_j = 1$$

$$\sum_{j=1}^{n} x_j = K$$

$$x_j \in \{0,1, \ldots, K\}$$

Note, x_j is the number of trials generating nominal value j. Obviously, the parameter $\Theta = (p_1, p_2, \ldots, p_n)^T$ does not include theoretical mean $E(X)$ and theoretical variance $V(X)$ but $E(X)$ and $V(X)$ of multinomial distribution can be calculated based on Θ as follows:

$$E(x_j) = Kp_j$$

$$V(x_j) = Kp_j(1 - p_j)$$ ■

When random variable X is considered as an observation, a statistic denoted $\tau(X)$ is function of X. For example, $\tau(X) = X$, $\tau(X) = aX + A$ where a is scalar constant and A is vector constant, and $\tau(X) = XX^T$ are statistics of X. Statistic $\tau(X)$ can be vector-by-vector functions, for example, $\tau(X) = (X, XX^T)^T$ is a very popular statistic of X.

In practice, if X is replaced by sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ including N observation X_i, a statistic is now function of X_i (s), for instance, quantities $\bar{X}$ and S defined below are statistics:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$S = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^T = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i^T\right) - \bar{X}\bar{X}^T$$

For multivariate normal distribution, $\bar{X}$ and S are estimates of theoretical mean μ and theoretical covariance matrix Σ. They are called sample mean and sample variance, respectively.

Statistic $\tau(X)$ is called sufficient statistic if it has all and only information to estimate parameter Θ. For example, sufficient statistic of normal PDF is $\tau(X) = (X, XX^T)^T$. In fact, parameter $\Theta = (\mu, \Sigma)^T$ of normal PDF, which includes theoretical mean μ and theoretical covariance matrix Σ, is totally determined based on all and only X and XX^T (there is no redundant information in $\tau(X)$) where X is observation considered as random variable, as follows:

$$\mu = E(X) = \int_X Xf(X|\Theta)\mathrm{d}X$$

$$\Sigma = E(X - \mu)(X - \mu)^T = E(XX^T) - \mu\mu^T$$

Similarly, given $X = (x_1, x_2, \ldots, x_n)^T$, sufficient statistic of multinomial PDF of K trials is $\tau(X) = (x_1, x_2, \ldots, x_n)^T$ due to:

$$p_j = \frac{E(x_j)}{K}, \forall j = \overline{1,n}$$

Given a sample containing observations, purpose of point estimation is to estimate unknown parameter Θ based on such sample. The result of estimation process is the estimate $\widehat{\Theta}$ as approximation of unknown Θ. Formula to calculate $\widehat{\Theta}$ based on sample is called estimator of Θ. As a convention, estimator of Θ is denoted $\widehat{\Theta}(X)$ or $\widehat{\Theta}(\mathcal{X})$ where X is an observation and $\mathcal{X}$ is sample including many observations. Actually, $\widehat{\Theta}(X)$ or $\widehat{\Theta}(\mathcal{X})$ is the same to $\widehat{\Theta}$ but the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

notation $\widehat{\Theta}(X)$ or $\widehat{\Theta}(\mathcal{X})$ implies that $\widehat{\Theta}$ is calculated based on observations. For example, given sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ including N observations iid X_i, estimator of theoretical mean μ of normal distribution is:

$$\hat{\mu} = \hat{\mu}(X) = \hat{\mu}(\mathcal{X}) = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

As usual, estimator of Θ is determined based on sufficient statistics which in turn are functions of observations where observations are considered as random variables. Estimation methods mentioned in this research are MLE, Maximum A Posteriori (MAP), and EM in which MAP and EM are variants of MLE.

According to viewpoint of Bayesian statistics, parameter Θ is also random variable. Equation 1.6 specifies Bayes' rule in which $f(\Theta|\xi)$ is called prior PDF (prior distribution) of Θ whereas $f(\Theta|X)$ is called posterior PDF (posterior distribution) of Θ given observation X. Note, ξ is parameter of the prior $f(\Theta|\xi)$, which is known as second-level parameter. For instance, if the prior $f(\Theta|\xi)$ is multivariate normal PDF, we have $\xi = (\mu_0, \Sigma_0^2)^T$ which are theoretical mean and theoretical covariance matrix of random variable Θ. Because ξ is constant, the prior PDF $f(\Theta|\xi)$ can be denoted $f(\Theta)$. Please pay attention that the posterior PDF $f(\Theta|X)$ is independent from ξ.

$$f(\Theta|X) = \frac{f(X|\Theta)f(\Theta|\xi)}{\int_\Theta f(X|\Theta)f(\Theta|\xi)} \tag{1.6}$$

In Bayes' rule, the PDF $f(X \mid \Theta)$ is called likelihood function. If posterior distribution $f(\Theta|X)$ has the same form of prior distribution $f(\Theta|\xi)$, such posterior distribution and prior distribution are called conjugate distributions (conjugate probabilities) and $f(\Theta|\xi)$ is called conjugate prior (Wikipedia, Conjugate prior, 2018) for likelihood function $f(X|\Theta)$. For example, if prior distribution $f(\Theta|\xi)$ is beta distribution and likelihood function $P(X|\Theta)$ follows binomial distribution then, posterior distribution $f(\Theta|X)$ is beta distribution too and hence, $f(\Theta|\xi)$ and $f(\Theta|X)$ are conjugate distributions. Shortly, whether posterior distribution and prior distribution are conjugate distributions depends on prior distribution and likelihood function. In some research, Θ is also called hypothesis.

When X is evaluated as observation, let $\widehat{\Theta}$ be estimate of Θ. It is calculated as a maximizer of the posterior PDF $f(\Theta|X)$ given X. Here data sample $\mathcal{X}$ has only one observation X as $\mathcal{X} = \{X\}$.

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, f(\Theta|X) = \underset{\Theta}{\operatorname{argmax}} \frac{f(X|\Theta)f(\Theta|\xi)}{\int_\Theta f(X|\Theta)f(\Theta|\xi)}$$

Because the prior PDF $f(\Theta|\xi)$ is assumed to be fixed and the value $\int_\Theta f(X|\Theta)f(\Theta|\xi)$ is constant with regard to Θ, we have:

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, f(\Theta|X) = \underset{\Theta}{\operatorname{argmax}}\, f(X|\Theta)$$

Obviously, MLE method determines $\widehat{\Theta}$ as a maximizer of the likelihood function $f(X \mid \Theta)$ with regard to Θ when X is evaluated as observation. It is interesting that the likelihood function $f(X|\Theta)$ is the PDF of X with parameter Θ. For convenience, MLE maximizes the natural logarithm of the likelihood function denoted $l(\Theta)$ instead of maximizing the likelihood function.

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, l(\Theta) = \underset{\Theta}{\operatorname{argmax}} \log\big(f(X|\Theta)\big) \tag{1.7}$$

Where $l(\Theta) = \log(f(X \mid \Theta))$ is called log-likelihood function of Θ. Recall that equation 1.7 implies the optimization problem. Note, $l(\Theta)$ is function of Θ if X is evaluated as observation.

$$l(\Theta) = l(\Theta|X) = \log\big(f(X|\Theta)\big) \tag{1.8}$$

Equation 1.7 is the simple result of MLE for estimating parameter based on observed sample. The notation $l(\Theta|X)$ implies that $l(\Theta)$ is determined based on X. If the log-likelihood function

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$l(\Theta)$ is first-order smooth function then, from equation 1.7, the estimate $\widehat{\Theta}$ can be solution of the equation created by setting the first-order derivative of $l(\Theta)$ regarding Θ to be zero. If solving such equation is too complex, some popular methods to solve optimization problem are Newton-Raphson (Burden & Faires, 2011, pp. 67-71), gradient descent (Ta, 2014), and Lagrange duality (Wikipedia, Karush–Kuhn–Tucker conditions, 2014).

For example, suppose $X = (x_1, x_2, \ldots, x_n)^T$ is vector and follows multivariate normal distribution,

$$f(X|\Theta) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Then the log-likelihood function is

$$l(\Theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)$$

Where μ and Σ are mean vector and covariance matrix of $f(X \mid \Theta)$, respectively with note that $\Theta = (\mu, \Sigma)^T$. The notation $|.|$ denotes determinant of given matrix and the notation Σ^{-1} denotes inverse of matrix Σ. Note, Σ is invertible and symmetric. Because normal PDF is smooth enough function, from equation 1.7, the estimate $\widehat{\Theta} = \left(\hat{\mu}, \hat{\Sigma}\right)^T$ is solution of the equation created by setting the first-order of $l(\Theta)$ regarding μ and Σ to be zero. The first-order partial derivative of $l(\Theta)$ with respect to μ is (Nguyen, 2015, p. 35):

$$\frac{\partial l(\Theta)}{\partial \mu} = (X-\mu)^T\Sigma^{-1}$$

Setting this partial derivative to be zero, we obtain:

$$(X-\mu)^T\Sigma^{-1} = 0 \Rightarrow X - \mu \Rightarrow \hat{\mu} = X$$

The first-order partial derivative of $l(\Theta)$ with respect to Σ is:

$$\frac{\partial l(\Theta)}{\partial \Sigma} = -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(X-\mu)(X-\mu)^T\Sigma^{-1}$$

Due to:

$$\frac{\partial \log(|\Sigma|)}{\partial \Sigma} = \Sigma^{-1}$$

And

$$\frac{\partial (X-\mu)^T\Sigma^{-1}(X-\mu)}{\partial \Sigma} = \frac{\partial \mathrm{tr}\left((X-\mu)(X-\mu)^T\Sigma^{-1}\right)}{\partial \Sigma}$$

Because Bilmes (Bilmes, 1998, p. 5) mentioned:

$$(X-\mu)^T\Sigma^{-1}(X-\mu) = \mathrm{tr}\left((X-\mu)(X-\mu)^T\Sigma^{-1}\right)$$

Where tr(A) is trace operator which takes sum of diagonal elements of square matrix, $\mathrm{tr}(A) = \sum_i a_{ii}$. This implies (Nguyen, 2015, p. 45):

$$\frac{\partial (X-\mu)^T\Sigma^{-1}(X-\mu)}{\partial \Sigma} = \frac{\partial \mathrm{tr}\left((X-\mu)(X-\mu)^T\Sigma^{-1}\right)}{\partial \Sigma} = -\Sigma^{-1}(X-\mu)(X-\mu)^T\Sigma^{-1}$$

Where Σ is symmetric and invertible matrix. Substituting the estimate $\hat{\mu}$ into the first-order partial derivative of $l(\Theta)$ with respect to Σ, we have:

$$\frac{\partial l(\Theta)}{\partial \Sigma} = -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(X-\hat{\mu})(X-\hat{\mu})^T\Sigma^{-1}$$

The estimate $\hat{\Sigma}$ is the solution of equation formed by setting the first-order partial derivative of $l(\Theta)$ regarding Σ to zero matrix. Let $(\mathbf{0})$ denote zero matrix.

$$(\mathbf{0}) = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

We have:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\frac{\partial l(\Theta)}{\partial \Sigma} = (\mathbf{0})$$
$$\Leftrightarrow -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(X-\hat{\mu})(X-\hat{\mu})^T\Sigma^{-1} = (\mathbf{0})$$
$$\Rightarrow -\Sigma + (X-\hat{\mu})(X-\hat{\mu})^T = (\mathbf{0})$$
$$\Rightarrow \hat{\Sigma} = (X-\hat{\mu})(X-\hat{\mu})^T$$

Finally, MLE results out the estimate $\widehat{\Theta}$ for normal distribution given observation X as follows:

$$\widehat{\Theta} = \left(\hat{\mu} = X, \hat{\Sigma} = (X-\hat{\mu})(X-\hat{\mu})^T\right)^T$$

When $\hat{\mu} = X$ then $\hat{\Sigma} = (\mathbf{0})$, which implies that the estimate $\hat{\Sigma}$ of covariance matrix is arbitrary with constraint that it is symmetric and invertible. This is reasonable because the sample is too small with only one observation X. When X is replaced by a sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ in which all X_i (s) are mutually independent and identically distributed (iid), it is easy to draw the following result by the similar way with equation 1.11.

$$\hat{\mu} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$
$$\hat{\Sigma} = S = \frac{1}{N}\sum_{i=1}^{N}(X_i-\hat{\mu})(X_i-\hat{\mu})^T = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i^T\right) - \hat{\mu}\hat{\mu}^T$$

Here, $\hat{\mu}$ and $\hat{\Sigma}$ are sample mean and sample variance ■

In practice, if X is observed as particular N observations $X_1, X_2, \ldots, X_N$. Let $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ be the observed sample of size N in which all X_i (s) are iid. The Bayes' rule specified by equation 1.6 is re-written as follows:

$$f(\Theta|\mathcal{X}) = \frac{f(\mathcal{X}|\Theta)f(\Theta|\xi)}{\int_\Theta f(\mathcal{X}|\Theta)f(\Theta|\xi)}$$

However, the meaning of Bayes' rule does not change. Because all X_i (s) are iid, the likelihood function becomes product of partial likelihood functions as follows:

$$f(\mathcal{X}|\Theta) = \prod_{i=1}^{N} f(X_i|\Theta) \quad (1.9)$$

The log-likelihood function of Θ becomes:

$$l(\Theta) = l(\Theta|\mathcal{X}) = \log\left(f(\mathcal{X}|\Theta)\right) = \log\left(\prod_{i=1}^{N} f(X_i|\Theta)\right) = \sum_{i=1}^{N}\log\left(f(X_i|\Theta)\right) \quad (1.10)$$

The notation $l(\Theta|\mathcal{X})$ implies that $l(\Theta)$ is determined based on $\mathcal{X}$. We have:

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, l(\Theta) = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{N}\log\left(f(X_i|\Theta)\right) \quad (1.11)$$

Equation 1.11 is the main result of MLE for estimating parameter based on observed sample. If the log-likelihood function $l(\Theta)$ is first-order smooth function then, from equation 1.11, the estimate $\widehat{\Theta}$ can be solution of the equation created by setting the first-order derivative of $l(\Theta)$ regarding Θ to be zero. If solving such equation is too complex, some popular methods to solve optimization problem are Newton-Raphson (Burden & Faires, 2011, pp. 67-71), gradient descent (Ta, 2014), and Lagrange duality (Wikipedia, Karush–Kuhn–Tucker conditions, 2014).

For example, suppose each $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$ is vector and follows multinomial distribution of K trials,

$$f(X_i|\Theta) = \frac{K!}{\prod_{j=1}^{n}(x_{ij}!)}\prod_{k=1}^{n} p_j^{x_{ij}}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

Where x_{ik} are integers and $\Theta = (p_1, p_2, \ldots, p_n)^T$ is the set of probabilities such that

$$\sum_{j=1}^{n} p_j = 1$$

$$\sum_{j=1}^{n} x_{ij} = K$$

$$x_{ij} \in \{0,1, \ldots, K\}$$

Note, x_{ik} is the number of trials generating nominal value k.

Given sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ in which all X_i (s) are iid, according to equation 1.10, the log-likelihood function is

$$l(\Theta) = l(\Theta|\mathcal{X}) = \sum_{i=1}^{n} \log\left(\frac{K!}{\prod_{j=1}^{n}(x_{ij}!)} \prod_{j=1}^{n} p_j^{x_{ij}}\right)$$

$$= \sum_{i=1}^{N}\left(\log(K!) - \sum_{j=1}^{n}\log(x_{ij}!) + \sum_{j=1}^{n} x_{ij}\log(p_j)\right)$$

Because there is the constraint $\sum_{j=1}^{n} p_j = 1$, we use Lagrange duality method to maximize $l(\Theta)$. The Lagrange function $la(\Theta, \lambda)$ is sum of $l(\Theta)$ and the constraint $\sum_{j=1}^{n} p_j = 1$ as follows:

$$la(\Theta,\lambda) = l(\Theta) + \lambda\left(1 - \sum_{j=1}^{n} p_j\right)$$

$$= \sum_{i=1}^{N}\left(\log(K!) - \sum_{j=1}^{n}\log(x_{ij}!) + \sum_{j=1}^{n} x_{ij}\log(p_j)\right) + \lambda\left(1 - \sum_{j=1}^{n} p_j\right)$$

Note, λ is called Lagrange multiplier. Of course, $la(\Theta, \lambda)$ is function of Θ and λ. Because multinomial PDF is smooth enough, the estimate $\widehat{\Theta} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)^T$ is solution of the equation created by setting the first-order of $la(\Theta)$ regarding p_j and λ to be zero. The first-order partial derivative of $la(\Theta)$ with respect to p_j is:

$$\frac{\partial la(\Theta)}{\partial p_j} = \frac{\sum_{i=1}^{N} x_{ij}}{p_j} - \lambda$$

Setting this partial derivative to be zero, we obtain following equation:

$$\frac{\sum_{i=1}^{N} x_{ij}}{p_j} - \lambda = 0 \Rightarrow \left(\sum_{i=1}^{N} x_{ij}\right) - \lambda p_j = 0$$

Summing this equation over n variables p_j, we obtain:

$$\sum_{j=1}^{n}\left(\left(\sum_{i=1}^{N} x_{ij}\right) - \lambda p_j\right) = \left(\sum_{i=1}^{N}\sum_{j=1}^{n} x_{ij}\right) - \lambda\sum_{j=1}^{n} p_j = 0$$

Due to

$$\sum_{j=1}^{n} p_j = 1$$

$$\sum_{j=1}^{n} x_{ij} = K$$

We have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$KN - \lambda = 0 \Rightarrow \lambda = KN$$

Substitute $\lambda = nN$ into equation

$$\left(\sum_{i=1}^{N} x_{ij}\right) - \lambda p_j = 0$$

We get the estimate $\widehat{\Theta} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)^T$ as follows:

$$\hat{p}_j = \frac{\sum_{i=1}^{N} x_{ij}}{KN} \blacksquare$$

Quality of estimation is measured by mean and variance of the estimate $\widehat{\Theta}$. The mean of $\widehat{\Theta}$ is:

$$E(\widehat{\Theta}) = \int_X \widehat{\Theta}(X)\, f(X|\Theta) \mathrm{d}X \tag{1.12}$$

The notation $\widehat{\Theta}(X)$ implies the formulation to calculate $\widehat{\Theta}$, which is resulted from MLE, MAP, or EM. Hence, $\widehat{\Theta}(X)$ is considered as function of X in the integral $\int_X \widehat{\Theta}(X)\, f(X|\Theta) \mathrm{d}X$. The $\widehat{\Theta}$ is unbiased estimate if $E(\widehat{\Theta}) = \Theta$. Otherwise, if $E(\widehat{\Theta}) \neq \Theta$ then, $\widehat{\Theta}$ is biased estimate. As usual, unbiased estimate is better than biased estimate. The condition $E(\widehat{\Theta}) = \Theta$ is the criterion to check if an estimate is unbiased, which is applied for all estimation methods.

The variance of $\widehat{\Theta}$ is:

$$V(\widehat{\Theta}) = \int_X \left(\widehat{\Theta}(X) - E(X)\right)\left(\widehat{\Theta}(X) - E(X)\right)^T f(X|\Theta) \mathrm{d}X \tag{1.13}$$

The smaller the variance $V(\widehat{\Theta})$, the better the $\widehat{\Theta}$ is.

For example, given multivariate normal distribution and given sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ where all X_i (s) are iid, the estimate $\widehat{\Theta} = (\hat{\mu}, \hat{\Sigma})^T$ from MLE is:

$$\hat{\mu} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (X_i - \hat{\mu})(X_i - \hat{\mu})^T$$

Due to:

$$E(\hat{\mu}) = E\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{N} E(X_i) = \frac{1}{N}\sum_{i=1}^{N} E(X) = \mu$$

Then $\hat{\mu}$ is unbiased estimate. We also have:

$$E(\hat{\Sigma}) = E\left(\frac{1}{N}\sum_{i=1}^{N} (X_i - \hat{\mu})(X_i - \hat{\mu})^T\right) = \frac{1}{N} E\left(\sum_{i=1}^{N} X_i X_i^T - \sum_{i=1}^{N} X_i \hat{\mu}^T - \sum_{i=1}^{N} \hat{\mu} X_i^T + \sum_{i=1}^{N} \hat{\mu}\hat{\mu}^T\right)$$

$$= \frac{1}{N} E\left(\sum_{i=1}^{N} X_i X_i^T - 2\sum_{i=1}^{N} \hat{\mu} X_i^T + \sum_{i=1}^{N} \hat{\mu}\hat{\mu}^T\right) = \frac{1}{N} E\left(\sum_{i=1}^{N} X_i X_i^T - 2\hat{\mu}\sum_{i=1}^{N} X_i^T + N\hat{\mu}\hat{\mu}^T\right)$$

(Due to $X_i \hat{\mu}^T = \hat{\mu} X_i^T$)

$$= \frac{1}{N} E\left(\sum_{i=1}^{N} X_i X_i^T - 2N\hat{\mu}\hat{\mu}^T + N\hat{\mu}\hat{\mu}^T\right) = \frac{1}{N} E\left(\sum_{i=1}^{N} X_i X_i^T - N\hat{\mu}\hat{\mu}^T\right)$$

(Due to $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$)

$$= \frac{1}{N}\sum_{i=1}^{N} E(X_i X_i^T) - E(\hat{\mu}\hat{\mu}^T) = \frac{1}{N}\sum_{i=1}^{N} E(XX^T) - E(\hat{\mu}\hat{\mu}^T) = E(XX^T) - E(\hat{\mu}\hat{\mu}^T)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

(Let X be random variable representing all iid X_i (s))

$= (\Sigma + \mu\mu^T) - (V(\hat{\mu}) + E(\hat{\mu})E(\hat{\mu})^T)$

(Due to $\Sigma = E(XX^T) - \mu\mu^T$ and the variance $V(\hat{\mu}) = E(\hat{\mu}\hat{\mu}^T) - E(\hat{\mu})E(\hat{\mu})^T$)

$= (\Sigma + \mu\mu^T) - (V(\hat{\mu}) + \mu\mu^T) = \Sigma - V(\hat{\mu})$

It is necessary to calculate the variance $V(\hat{\mu})$. In fact, we have:

$$V(\hat{\mu}) = V\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} V(X_i) = \frac{1}{N^2}\sum_{i=1}^{N} V(X) = \frac{1}{N}V(X) = \frac{1}{N}\Sigma$$

Therefore, we have:

$$E(\hat{\Sigma}) = \Sigma - \frac{1}{N}\Sigma = \frac{N-1}{N}\Sigma$$

Hence, we conclude that $\hat{\Sigma}$ is biased estimate because of $E(\hat{\Sigma}) \neq \Sigma$ ■

Without loss of generality, suppose parameter Θ is vector, the second-order derivative of the log-likelihood function $l(\Theta)$ is called likelihood Hessian matrix (Zivot, 2009, p. 7) denoted $S(\Theta)$.

$$S(\Theta) = S(\Theta|X) = D^2 l(\Theta|X) \tag{1.14}$$

Suppose $\Theta = (\theta_1, \theta_2, \ldots, \theta_r)^T$ where there are r partial parameters θ_k, equation 1.14 is expended as follows:

$$D^2 l(\Theta|X) = \frac{\mathrm{d}^2 l(\Theta|X)}{\mathrm{d}\Theta^2} = \begin{pmatrix} \frac{\partial^2 l(\Theta|X)}{\partial\theta_1^2} & \frac{\partial^2 l(\Theta|X)}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 l(\Theta|X)}{\partial\theta_1\partial\theta_r} \\ \frac{\partial^2 l(\Theta|X)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 l(\Theta|X)}{\partial\theta_2^2} & \cdots & \frac{\partial^2 l(\Theta|X)}{\partial\theta_2\partial\theta_r} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\Theta|X)}{\partial\theta_r\partial\theta_1} & \frac{\partial^2 l(\Theta|X)}{\partial\theta_r\partial\theta_2} & \cdots & \frac{\partial^2 l(\Theta|X)}{\partial\theta_r^2} \end{pmatrix}$$

Where,

$$\frac{\partial^2 l(\Theta|X)}{\partial\theta_i\partial\theta_j} = \frac{\partial}{\partial\theta_i}\left(\frac{\partial l(\Theta|X)}{\partial\theta_j}\right)$$

$$\frac{\partial^2 l(\Theta|X)}{\partial\theta_i^2} = \frac{\partial^2 l(\Theta|X)}{\partial\theta_i\partial\theta_i}$$

The notation $l(\Theta|X)$ implies that $l(\Theta)$ is determined based on X, according to equation 1.8. The notation $S(\Theta|X)$ implies $S(\Theta)$ is calculated based on X. If sample $\mathcal{X}$ replaces X then,

$$S(\Theta) = S(\Theta|\mathcal{X}) = D^2 l(\Theta|\mathcal{X}) \tag{1.15}$$

Where $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ be the observed sample of size N in which all X_i (s) are iid. The notation $l(\Theta|\mathcal{X})$ implies that $l(\Theta)$ is determined based on $\mathcal{X}$, according to equation 1.11. The notation $S(\Theta|\mathcal{X})$ implies $S(\Theta)$ is calculated based on $\mathcal{X}$.

The negative expectation of likelihood Hessian matrix is called information matrix or Fisher information matrix denoted $I(\Theta)$.

$$I(\Theta) = -E(S(\Theta)) \tag{1.16}$$

If $S(\Theta)$ is calculated by equation 1.14 with observation X then, $I(\Theta)$ becomes:

$$I(\Theta) = I(\Theta|X) = -E(S(\Theta|X)) = -\int_X D^2 l(\Theta|X) f(X|\Theta)\mathrm{d}X \tag{1.17}$$

The notation $l(\Theta|X)$ implies that $l(\Theta)$ is determined based on X, according to equation 1.8. The notation $I(\Theta|X)$ implies $I(\Theta)$ is calculated based on X. Note, $D^2 l(\Theta|X)$ is considered as function of X in the integral $\int_X D^2 l(\Theta|X) f(X|\Theta)\mathrm{d}X$.

If $S(\Theta)$ is calculated by equation 1.15 with observation sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ in which all X_i (s) are iid then, $I(\Theta)$ becomes:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$I(\Theta) = I(\Theta|\mathcal{X}) = -E\big(S(\Theta|\mathcal{X})\big) = N * I(\Theta|X) = -N\int_X D^2 l(\Theta|X) f(X|\Theta)\mathrm{d}X \quad (1.18)$$

Where X is random variable representing every X_i. The notation $I(\Theta|\mathcal{X})$ implies $I(\Theta)$ is calculated based on $\mathcal{X}$. Note, $D^2 l(\Theta|X)$ is considered as function of X in the integral $\int_X D^2 l(\Theta|X) f(X|\Theta)\mathrm{d}X$. Following is proof of equation 1.18.

$I(\Theta) = I(\Theta|\mathcal{X}) = -E\big(S(\Theta|\mathcal{X})\big) = -E\big(D^2 l(\Theta|\mathcal{X})\big)$

(The notation $l(\Theta|\mathcal{X})$ implies that $l(\Theta)$ is determined based on $\mathcal{X}$)

$$= -E\left(\sum_{i=1}^{N} D^2 l(\Theta|X_i)\right)$$

(Due to equation 1.8 and iid X_i (s))

$$= -\sum_{i=1}^{N} E\big(D^2 l(\Theta|X_i)\big) = -\sum_{i=1}^{N}\int_X D^2 l(\Theta|X_i) f(X_i|\Theta)\mathrm{d}X_i$$

$$= -\sum_{i=1}^{N}\int_X D^2 l(\Theta|X) f(X|\Theta)\mathrm{d}X$$

(Let X be random variable representing every X_i)

$$= -N\int_X D^2 l(\Theta|X) f(X|\Theta)\mathrm{d}X = N * I(\Theta|X)\blacksquare$$

For MLE method, the inverse of estimator information matrix is called Cramer-Rao lower bound denoted $CR\big(\widehat{\Theta}\big)$.

$$CR\big(\widehat{\Theta}\big) = I(\Theta)^{-1} \quad (1.19)$$

Where $I(\Theta)$ is calculated by equation 1.17 or equation 1.18. Any covariance matrix of a MLE estimate $\widehat{\Theta}$ has such Cramer-Rao lower bound. Such Cramer-Rao lower bound becomes $V\big(\widehat{\Theta}\big)$ if and only if $\widehat{\Theta}$ is unbiased, (Zivot, 2009, p. 11):

$$\begin{aligned} V\big(\widehat{\Theta}\big) &\geq CR\big(\widehat{\Theta}\big) \text{ if } \widehat{\Theta} \text{ biased} \\ V\big(\widehat{\Theta}\big) &= CR\big(\widehat{\Theta}\big) \text{ if } \widehat{\Theta} \text{ unbiased} \end{aligned} \quad (1.20)$$

Note, equation 1.19 and equation 1.20 are only valid for MLE method. The sign "≥" implies lower bound. In other words, Cramer-Rao lower bound is variance of the optimal MLE estimate. Moreover, beside the criterion $E\big(\widehat{\Theta}\big) = \Theta$, equation 1.20 can be used as another criterion to check if an estimate is unbiased. However, the criterion $E\big(\widehat{\Theta}\big) = \Theta$ is applied for all estimation methods whereas equation 1.20 is only applied for MLE.

Suppose $\Theta = (\theta_1, \theta_2, \ldots, \theta_r)^T$ where there are r partial parameter θ_k, so the estimate is $\widehat{\Theta} = \big(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r\big)^T$. Each element on diagonal of the Cramer-Rao lower bound is lower bound of a variance of $\hat{\theta}_k$, denoted $V\big(\hat{\theta}_k\big)$. Let $CR\big(\hat{\theta}_k\big)$ be lower bound of $V\big(\hat{\theta}_k\big)$, of course we have:

$$\begin{aligned} V\big(\hat{\theta}_k\big) &\geq CR\big(\hat{\theta}_k\big) \text{ if } \hat{\theta}_k \text{ biased} \\ V\big(\hat{\theta}_k\big) &= CR\big(\hat{\theta}_k\big) \text{ if } \hat{\theta}_k \text{ unbiased} \end{aligned} \quad (1.21)$$

The sign "≥" implies lower bound. Derived from equation 1.18 and equation 1.19, $CR\big(\hat{\theta}_k\big)$ is specified by equation 1.22.

$$\begin{aligned} I\big(\hat{\theta}_k\big) &= -N * E\left(\frac{\partial^2 l(\Theta|X)}{\partial\theta_k^2}\right) = -N\int_X \frac{\partial^2 l(\Theta|X)}{\partial\theta_k^2} f(X|\Theta)\mathrm{d}X \\ CR\big(\hat{\theta}_k\big) &= I\big(\hat{\theta}_k\big)^{-1} = -\frac{1}{N}\left(\int_X \frac{\partial^2 l(\Theta|X)}{\partial\theta_k^2} f(X|\Theta)\mathrm{d}X\right)^{-1} \end{aligned} \quad (1.22)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

Where N is size of sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ in which all X_i (s) are iid. If there is only one observation X then, $N = 1$. Of course, $I(\hat{\theta}_k)$ is information matrix of $\hat{\theta}_k$. If $\hat{\theta}_k$ is univariate, $I(\hat{\theta}_k)$ is scalar, which called information value.

For example, let $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ be the observed sample of size N with note that all X_i (s) are iid, given multivariate normal PDF as follows:

$$f(X|\Theta) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Where n is dimension of vector X and $\Theta = (\mu, \Sigma)^T$ with note that μ is theoretical mean vector and Σ is theoretical covariance matrix. Note, Σ is invertible and symmetric. From previous example, the MLE estimate $\widehat{\Theta} = \left(\hat{\mu}, \hat{\Sigma}\right)^T$ given $\mathcal{X}$ is:

$$\hat{\mu} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu})(X_i - \hat{\mu})^T$$

Mean and variance of $\hat{\mu}$ from previous example are:

$$E(\hat{\mu}) = \mu$$

$$V(\hat{\mu}) = \frac{1}{N}\Sigma$$

We knew that $\hat{\mu}$ is unbiased estimate with criterion $E(\hat{\mu}) = \mu$. Now we check again if $\hat{\mu}$ is unbiased estimate with equation 1.21 as another criterion for MLE. Hence, we firstly calculate the lower bound $CR(\hat{\mu})$ and then compare it with the variance $V(\hat{\mu})$. In fact, according to equation 1.8, the log-likelihood function is:

$$l(\Theta|X) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)$$

The partial first-order derivative of $l(\Theta|X)$ with regard to μ is (Nguyen, 2015, p. 35):

$$\frac{\partial l(\Theta|X)}{\partial \mu} = (X-\mu)^T\Sigma^{-1}$$

$$\left(\text{due to } \frac{\partial (X-\mu)^T\Sigma^{-1}(X-\mu)}{\partial \mu} = -2(X-\mu)^T\Sigma^{-1} \text{ when } \Sigma \text{ is symmetric}\right)$$

The partial second-order derivative of $l(\Theta|X)$ with regard to μ is (Nguyen, 2015, p. 36):

$$\frac{\partial^2 l(\Theta|X)}{\partial \mu^2} = \frac{\partial}{\partial \mu}\left(\frac{\partial l(\Theta|X)}{\partial \mu}\right) = \frac{\partial}{\partial \mu}((X-\mu)^T\Sigma^{-1}) = -(\Sigma^{-1})^T = \Sigma^{-1}$$

(Due to Σ is symmetric)

According to equation 1.22, the lower bound $CR(\hat{\mu})$ is:

$$CR(\hat{\mu}) = -\frac{1}{N}\left(\int_X \frac{\partial^2 l(\Theta|X)}{\partial \mu^2} f(X|\Theta)\mathrm{d}X\right)^{-1} = \frac{1}{N}\left(\int_X \Sigma^{-1} f(X|\Theta)\mathrm{d}X\right)^{-1}$$

$$= \frac{1}{N}\left(\Sigma^{-1}\int_X f(X|\Theta)\mathrm{d}X\right)^{-1} = \frac{1}{N}\Sigma = V(\hat{\mu})$$

Due to $V(\hat{\mu}) = CR(\hat{\mu})$, $\hat{\mu}$ is unbiased estimate according to the criterion specified by equation 1.21.

Mean of $\hat{\Sigma}$ from previous example is:

$$E\left(\hat{\Sigma}\right) = \frac{N-1}{N}\Sigma$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

We knew that $\hat{\Sigma}$ is biased estimate because $E(\hat{\Sigma}) \neq \Sigma$. Now we check again if $\hat{\Sigma}$ is biased estimate with equation 1.21 as another criterion for MLE. The partial first-order derivative of $l(\Theta|X)$ with regard to Σ is:

$$\frac{\partial l(\Theta|X)}{\partial \Sigma} = -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(X-\mu)(X-\mu)^T\Sigma^{-1}$$

Due to:

$$\frac{\partial \log(|\Sigma|)}{\partial \Sigma} = \Sigma^{-1}$$

And

$$\frac{\partial (X-\mu)^T\Sigma^{-1}(X-\mu)}{\partial \Sigma} = \frac{\partial \mathrm{tr}\left((X-\mu)(X-\mu)^T\Sigma^{-1}\right)}{\partial \Sigma}$$

Because Bilmes (Bilmes, 1998, p. 5) mentioned:

$$(X-\mu)^T\Sigma^{-1}(X-\mu) = \mathrm{tr}\left((X-\mu)(X-\mu)^T\Sigma^{-1}\right)$$

Where tr(A) is trace operator which takes sum of diagonal elements of matrix $\mathrm{tr}(A) = \sum_i a_{ii}$.

This implies (Nguyen, 2015, p. 45):

$$\frac{\partial (X-\mu)^T\Sigma^{-1}(X-\mu)}{\partial \Sigma} = \frac{\partial \mathrm{tr}\left((X-\mu)(X-\mu)^T\Sigma^{-1}\right)}{\partial \Sigma} = -\Sigma^{-1}(X-\mu)(X-\mu)^T\Sigma^{-1}$$

According to equation 1.22, the lower bound $CR(\hat{\Sigma})$ is:

$$CR(\hat{\Sigma}) = -\frac{1}{N}\left(\int_X \frac{\partial^2 l(\Theta|X)}{\partial \Sigma^2} f(X|\Theta)\mathrm{d}X\right)^{-1} = -\frac{1}{N}\left(\int_X \frac{\partial}{\partial \Sigma}\left(\frac{\partial l(\Theta|X)}{\partial \Sigma}\right) f(X|\Theta)\mathrm{d}X\right)^{-1}$$

$$= -\frac{1}{N}\left(\frac{\partial}{\partial \Sigma}\left(\int_X \frac{\partial l(\Theta|X)}{\partial \Sigma} f(X|\Theta)\mathrm{d}X\right)\right)^{-1}$$

(Due to $l(\Theta|X)$ is smooth enough)

$$= -\frac{1}{N}\left(\frac{\partial}{\partial \Sigma}\left(\int_X \left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(X-\mu)(X-\mu)^T\Sigma^{-1}\right) f(X|\Theta)\mathrm{d}X\right)\right)^{-1}$$

$$= -\frac{1}{N}\left(\frac{\partial}{\partial \Sigma}\left(-\frac{1}{2}\Sigma^{-1}\int_X f(X|\Theta)\mathrm{d}X + \frac{1}{2}\int_X \Sigma^{-1}(X-\mu)(X-\mu)^T\Sigma^{-1} f(X|\Theta)\mathrm{d}X\right)\right)^{-1}$$

$$= -\frac{1}{N}\left(\frac{\partial}{\partial \Sigma}\left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\int_X \Sigma^{-1}(X-\mu)(X-\mu)^T\Sigma^{-1} f(X|\Theta)\mathrm{d}X\right)\right)^{-1}$$

$$= -\frac{1}{N}\left(\frac{\partial}{\partial \Sigma}\left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\Sigma^{-1}\int_X (X-\mu)(X-\mu)^T f(X|\Theta)\mathrm{d}X\right)\right)^{-1}$$

(Because Σ^{-1} and $(X-\mu)(X-\mu)^T$ are symmetric matrices)

$$= -\frac{1}{N}\left(\frac{\partial}{\partial \Sigma}\left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\Sigma^{-1}\Sigma\right)\right)^{-1} = -\frac{1}{N}\left(\frac{\partial}{\partial \Sigma}\left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\right)\right)^{-1}$$

$$= -\frac{1}{N}\left(\frac{\partial}{\partial \Sigma}((\mathbf{0}))\right)^{-1}$$

Where ($\mathbf{0}$) is zero matrix. This implies the lower bound $CR(\hat{\Sigma})$ is inexistent. Hence, $\hat{\Sigma}$ is biased estimate. Even there is no unbiased estimate of variance for normal distribution by MLE ■

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

MLE ignores prior PDF $f(\Theta|\xi)$ because $f(\Theta|\xi)$ is assumed to be fixed but Maximum A Posteriori (MAP) method (Wikipedia, Maximum a posteriori estimation, 2017) concerns $f(\Theta|\xi)$ in maximization task when $\int_\Theta f(X|\Theta)f(\Theta|\xi)$ is constant with regard to Θ.

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, f(\Theta|X) = \underset{\Theta}{\operatorname{argmax}} \frac{f(X|\Theta)f(\Theta|\xi)}{\int_\Theta f(X|\Theta)f(\Theta|\xi)} = \underset{\Theta}{\operatorname{argmax}}\, f(X|\Theta)f(\Theta|\xi)$$

Let $f(X, \Theta \mid \xi)$ be the joint PDF of X and Θ where Θ is also random variable too. Note, ξ is parameter in the prior PDF $f(\Theta|\xi)$. The likelihood function in MAP is also $f(X, \Theta \mid \xi)$.

$$f(X, \Theta|\xi) = f(X|\Theta)f(\Theta|\xi) \tag{1.23}$$

Theoretical mean and variance of X are based on the joint PDF $f(X, \Theta \mid \xi)$ as follows:

$$E(X) = \int_X \int_\Theta X f(X, \Theta|\xi) \mathrm{d}X \mathrm{d}\Theta \tag{1.24}$$

$$V(X) = \int_X \int_\Theta \big(X - E(X)\big)\big(X - E(X)\big)^T f(X, \Theta|\xi) \mathrm{d}X \mathrm{d}\Theta \tag{1.25}$$

Theoretical mean and variance of Θ are based on $f(\Theta|\xi)$ because $f(\Theta|\xi)$ is function of only Θ when ξ is constant.

$$E(\Theta) = \int_X \int_\Theta \Theta f(X, \Theta|\xi) \mathrm{d}X \mathrm{d}\Theta = \int_\Theta \Theta f(\Theta|\xi) \mathrm{d}\Theta \tag{1.26}$$

$$\begin{aligned} V(\Theta) &= \int_X \int_\Theta \big(\Theta - E(\Theta)\big)\big(\Theta - E(\Theta)\big)^T f(X, \Theta|\xi) \mathrm{d}X \mathrm{d}\Theta \\ &= \int_\Theta \big(\Theta - E(\Theta)\big)\big(\Theta - E(\Theta)\big)^T f(\Theta|\xi) \mathrm{d}\Theta \\ &= E(\Theta\Theta^T|\xi) - E(\Theta|\xi)E(\Theta^T|\xi) \end{aligned} \tag{1.27}$$

In general, statistics of Θ are still based on $f(\Theta|\xi)$. Given sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ in which all X_i (s) are iid, the likelihood function becomes:

$$f(\mathcal{X}, \Theta|\xi) = \prod_{i=1}^{N} f(X_i, \Theta|\xi) \tag{1.28}$$

The log-likelihood function $\ell(\Theta)$ in MAP is re-defined with observation X or sample $\mathcal{X}$ as follows:

$$\ell(\Theta) = \log\big(f(X, \Theta|\xi)\big) = l(\Theta) + \log\big(f(\Theta|\xi)\big) \tag{1.29}$$

$$\ell(\Theta) = \log\big(f(\mathcal{X}, \Theta|\xi)\big) = l(\Theta) + \log\big(f(\Theta|\xi)\big) \tag{1.30}$$

Where $l(\Theta)$ is specified by equation 1.8 with observation X or equation 1.10 with sample $\mathcal{X}$. Therefore, the estimate $\widehat{\Theta}$ is determined according to MAP as follows:

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\big(\ell(\Theta)\big) = \underset{\Theta}{\operatorname{argmax}}\Big(l(\Theta) + \log\big(f(\Theta|\xi)\big)\Big) \tag{1.31}$$

Good information provided by the prior $f(\Theta|\xi)$ can improve quality of estimation. Essentially, MAP is an improved variant of MLE. Later on, we also recognize that EM algorithm is also a variant of MLE. All of them aim to maximize log-likelihood functions. Likelihood Hessian matrix $S\big(\widehat{\Theta}\big)$, information matrix $I\big(\widehat{\Theta}\big)$, and Cramer-Rao lower bound $CR\big(\widehat{\Theta}\big)$, $CR\big(\hat{\theta}_k\big)$ are extended in MAP with the new likelihood function $\ell(\Theta)$.

$$S(\Theta) = D^2\ell(\Theta)$$
$$I(\Theta) = -E\big(S(\Theta)\big)$$
$$CR\big(\widehat{\Theta}\big) = I(\Theta)^{-1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$I(\hat{\theta}_k) = -N * E\left(\frac{\partial^2 \ell(\Theta)}{\partial \theta_k^2}\right) = -N \int_X \int_\Theta \frac{\partial^2 \ell(\Theta)}{\partial \theta_k^2} f(X, \Theta|\xi) \mathrm{d}X \mathrm{d}\Theta$$

$$CR(\hat{\theta}_k) = I(\hat{\theta}_k)^{-1}$$

Where N is size of sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ in which all X_i (s) are iid. If there is only one observation X then, $N = 1$.

Mean and variance of the estimate $\widehat{\Theta}$ which are used to measure estimation quality are not changed except that the joint PDF $f(X, \Theta \mid \xi)$ is used instead.

$$E(\widehat{\Theta}) = \int_X \int_\Theta \widehat{\Theta}(X, \Theta) f(X, \Theta|\xi) \mathrm{d}X \mathrm{d}\Theta \tag{1.32}$$

$$V(\widehat{\Theta}) = \int_X \int_\Theta \left(\widehat{\Theta}(X, \Theta) - E(X)\right)\left(\widehat{\Theta}(X, \Theta) - E(X)\right)^T f(X, \Theta|\xi) \mathrm{d}X \mathrm{d}\Theta \tag{1.33}$$

The notation $\widehat{\Theta}(X, \Theta)$ implies the formulation to calculate $\widehat{\Theta}$, which is considered as function of X and Θ in the integral $\int_X \int_\Theta \widehat{\Theta}(X, \Theta) f(X, \Theta|\xi) \mathrm{d}X \mathrm{d}\Theta$. Recall the $\widehat{\Theta}$ is unbiased estimate if $E(\widehat{\Theta}) = \Theta$. Otherwise, if $E(\widehat{\Theta}) \neq \Theta$ then, $\widehat{\Theta}$ is biased estimate. Moreover, the smaller the variance $V(\widehat{\Theta})$, the better the $\widehat{\Theta}$ is. Recall that there are two criteria to check if $\widehat{\Theta}$ is unbiased estimate. Concretely, $\widehat{\Theta}$ is unbiased estimate if one of two following conditions is satisfied:

$$E(\widehat{\Theta}) = \Theta$$

$$V(\widehat{\Theta}) = CR(\widehat{\Theta})$$

The criterion $V(\widehat{\Theta}) = CR(\widehat{\Theta})$ is expended for MAP.

It is necessary to have an example for parameter estimation with MAP. Given sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ in which all X_i (s) are iid. Each n-dimension X_i has following multivariate normal PDF:

$$f(X_i|\Theta) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X_i - \mu)^T \Sigma^{-1} (X_i - \mu)\right)$$

Where μ and Σ are mean vector and covariance matrix of $f(X \mid \Theta)$, respectively with note that $\Theta = (\mu, \Sigma)^T$. The notation $|.|$ denotes determinant of given matrix and the notation Σ^{-1} denotes inverse of matrix Σ. Note, Σ is invertible and symmetric.

In $\Theta = (\mu, \Sigma)^T$, suppose only μ distributes normally with parameter $\xi = (\mu_0, \Sigma_0)$ where μ_0 and Σ_0 are theoretical mean and covariance matrix of μ. Thus, Σ is variable but not random variable. The second-level parameter ξ is constant. The prior PDF $f(\Theta|\xi)$ becomes $f(\mu|\xi)$, which specified as follows:

$$f(\Theta|\xi) = f(\mu|\mu_0, \Sigma_0) = (2\pi)^{-\frac{n}{2}} |\Sigma_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mu - \mu_0)^T \Sigma_0^{-1} (\mu - \mu_0)\right)$$

Note, μ_0 is n-element vector like μ and Σ_0 is $n_x n$ matrix like Σ. Of course, Σ_0 is also invertible and symmetric. Suppose $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T$, $\mu_0 = (\mu_{01}, \mu_{02}, \ldots, \mu_{0n})^T$, and

$$\Sigma_0 = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{pmatrix}$$

It is deduced that μ_{0j} is theoretical mean of μ_j whereas δ_{ij} ($i \neq j$) is covariance of μ_i and μ_j. Especially, δ_{ii} is variance of μ_i.

Theoretical mean of X is:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$E(X) = \int_X \int_\Theta X f(X,\Theta|\xi)\mathrm{d}X\mathrm{d}\Theta = \int_X \int_\Theta X f(X|\Theta) f(\Theta|\xi)\mathrm{d}X\mathrm{d}\Theta$$
$$= \int_\Theta \left(\int_\mathrm{X} X f(X|\Theta)\mathrm{d}X \right) f(\Theta|\xi)\mathrm{d}\Theta = \int_\Theta \mu f(\Theta|\xi)\mathrm{d}\Theta = \int_\mu \mu f(\mu|\mu_0,\Sigma_0)\mathrm{d}\mu$$
$$= E(\mu) = \mu_0$$

Theoretical variance of X is:

$$V(X) = \int_X \int_\Theta \big(X - E(X)\big)\big(X - E(X)\big)^T f(X,\Theta|\xi)\mathrm{d}X\mathrm{d}\Theta$$
$$= \int_X \int_\Theta \big(X - E(X)\big)\big(X - E(X)\big)^T f(X|\Theta) f(\Theta|\xi)\mathrm{d}X\mathrm{d}\Theta$$
$$= \int_\Theta \left(\int_\mathrm{X} \big(X - E(X)\big)\big(X - E(X)\big)^T f(X|\Theta)\mathrm{d}X \right) f(\Theta|\xi)\mathrm{d}\Theta$$
$$= \int_\Theta \Sigma f(\Theta|\xi)\mathrm{d}\Theta = \int_\mu \Sigma f(\mu|\mu_0,\Sigma_0)\mathrm{d}\mu = \Sigma$$

The log-likelihood function in MAP is

$$\ell(\Theta) = \log\big(f(\mu|\xi)\big) + l(\Theta) = \log\big(f(\mu|\xi)\big) + \sum_{i=1}^{N} \log\big(f(X_i|\Theta)\big)$$
$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)$$
$$+ \sum_{i=1}^{N}\left(-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) \right)$$

Because normal PDF is smooth enough, from equation 1.24, the estimate $\widehat{\Theta} = \big(\hat{\mu},\hat{\Sigma}\big)^T$ is solution of the equation created by setting the first-order of $\ell(\Theta)$ regarding μ and Σ to be zero. Due to (Nguyen, 2015, p. 35):

$$\frac{\partial}{\partial\mu}\big((X-\mu)^T\Sigma^{-1}(X-\mu)\big) = -2(X-\mu)^T\Sigma^{-1}$$

And (Nguyen, 2015, p. 35)

$$\frac{\partial}{\partial\mu}((\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)) = (\mu-\mu_0)^T(\Sigma_0^{-1} + (\Sigma_0^{-1})^T) = (\mu-\mu_0)^T(\Sigma_0^{-1} + \Sigma_0^{-1})$$
$$= 2(\mu-\mu_0)^T\Sigma_0^{-1}$$

The first-order partial derivative of $\ell(\Theta)$ with respect to μ is:

$$\frac{\partial\ell(\Theta)}{\partial\mu} = -(\mu-\mu_0)^T\Sigma_0^{-1} + \sum_{i=1}^{N}(X_i-\mu)^T\Sigma^{-1}$$
$$= -\mu^T\Sigma_0^{-1} + \mu_0^T\Sigma_0^{-1} + \left(\sum_{i=1}^{N} X_i^T\right)\Sigma^{-1} - N\mu^T\Sigma^{-1}$$
$$= -\mu^T(\Sigma_0^{-1} + N\Sigma^{-1}) + \mu_0^T\Sigma_0^{-1} + \left(\sum_{i=1}^{N} X_i^T\right)\Sigma^{-1}$$

Setting this partial derivative to be zero, we obtain:

$$-\mu^T(\Sigma_0^{-1} + N\Sigma^{-1}) + \mu_0^T\Sigma_0^{-1} + \left(\sum_{i=1}^{N} X_i^T\right)\Sigma^{-1} = 0$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\Rightarrow (\Sigma_0^{-1} + N\Sigma^{-1})^T \mu = \Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{i=1}^{N} X_i$$

$$\Rightarrow (\Sigma_0^{-1} + N\Sigma^{-1})\mu = \Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{i=1}^{N} X_i$$

$$\Rightarrow (\Sigma\Sigma_0^{-1} + NI)\mu = \Sigma\Sigma_0^{-1}\mu_0 + \sum_{i=1}^{N} X_i$$

Where I is identity matrix. Let,

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

We obtain the following equation to estimate μ and Σ:

$$\mu = (\Sigma\Sigma_0^{-1} + NI)^{-1}(\Sigma\Sigma_0^{-1}\mu_0 + N\bar{X})$$

The first-order partial derivative of $l(\Theta)$ with respect to Σ is:

$$\frac{\partial \ell(\Theta)}{\partial \Sigma} = \sum_{i=1}^{N}\left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(X_i - \mu)(X_i - \mu)^T\Sigma^{-1}\right)$$

Due to:

$$\frac{\partial \log(|\Sigma|)}{\partial \Sigma} = \Sigma^{-1}$$

And

$$\frac{\partial (X_i - \mu)^T\Sigma^{-1}(X_i - \mu)}{\partial \Sigma} = \frac{\partial \mathrm{tr}\left((X_i - \mu)(X_i - \mu)^T\Sigma^{-1}\right)}{\partial \Sigma}$$

Because Bilmes (Bilmes, 1998, p. 5) mentioned:

$$(X_i - \mu)^T\Sigma^{-1}(X_i - \mu) = \mathrm{tr}\left((X_i - \mu)(X_i - \mu)^T\Sigma^{-1}\right)$$

Where tr(A) is trace operator which takes sum of diagonal elements of square matrix, $\mathrm{tr}(A) = \sum_i a_{ii}$. This implies (Nguyen, 2015, p. 45):

$$\frac{\partial (X_i - \mu)^T\Sigma^{-1}(X_i - \mu)}{\partial \Sigma} = \frac{\partial \mathrm{tr}\left((X_i - \mu)(X_i - \mu)^T\Sigma^{-1}\right)}{\partial \Sigma} = -\Sigma^{-1}(X_i - \mu)(X_i - \mu)^T\Sigma^{-1}$$

Where Σ is symmetric and invertible matrix. The estimate $\hat{\Sigma}$ is the solution of equation formed by setting the first-order partial derivative of $l(\Theta)$ regarding Σ to zero matrix. Let ($\mathbf{0}$) denote zero matrix.

$$(\mathbf{0}) = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

We have:

$$\frac{\partial \ell(\Theta)}{\partial \Sigma} = (\mathbf{0})$$

$$\Leftrightarrow \sum_{i=1}^{N}\left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(X_i - \mu)(X_i - \mu)^T\Sigma^{-1}\right) = (\mathbf{0})$$

$$\Rightarrow \sum_{i=1}^{N}(-\Sigma + (X_i - \mu)(X_i - \mu)^T) = (\mathbf{0})$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\Rightarrow \Sigma = \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)(X_i-\mu)^T = \frac{1}{N}\sum_{i=1}^{N}(X_iX_i^T - X_i\mu^T - \mu X_i^T + \mu\mu^T)$$

$$= \frac{1}{N}\sum_{i=1}^{N}(X_iX_i^T - \mu X_i^T - \mu X_i^T + \mu\mu^T)$$

$$\Rightarrow \Sigma = \left(\frac{1}{N}\sum_{i=1}^{N}X_iX_i^T\right) - \frac{2}{N}\mu\sum_{i=1}^{N}X_i^T + \mu\mu^T = \left(\frac{1}{N}\sum_{i=1}^{N}X_iX_i^T\right) - 2\mu\bar{X} + \mu\mu^T$$

MAP results out a system of two equations whose solution is the estimate $\widehat{\Theta} = (\hat{\mu}, \hat{\Sigma})^T$ as follows:

$$\begin{cases} \mu = (\Sigma\Sigma_0^{-1} + NI)^{-1}(\Sigma\Sigma_0^{-1}\mu_0 + N\bar{X}) \\ \Sigma = \left(\frac{1}{N}\sum_{i=1}^{N}X_iX_i^T\right) - 2\mu\bar{X} + \mu\mu^T \end{cases}$$

Where I is identity matrix and

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N}X_i$$

Because Σ is independent from the prior PDF $f(\mu \mid \mu_0, \Sigma_0)$, it is estimated by MLE as usual,

$$\hat{\Sigma} = \left(\frac{1}{N}\sum_{i=1}^{N}X_iX_i^T\right) - \bar{X}\bar{X}^T$$

The estimate $\hat{\Sigma}$ in MAP here is as same as the one in MLE and so, it is biased. Substituting $\hat{\Sigma}$ for Σ, we obtain the estimate $\hat{\mu}$ in MAP:

$$\hat{\mu} = \left(\hat{\Sigma}\Sigma_0^{-1} + NI\right)^{-1}\left(\hat{\Sigma}\Sigma_0^{-1}\mu_0 + N\bar{X}\right)$$

Note,

$$E(\bar{X}) = E\left(\frac{1}{N}\sum_{i=1}^{N}X_i\right) = \frac{1}{N}\sum_{i=1}^{N}E(X_i) = E(X) = \mu_0$$

$$V(\bar{X}) = V\left(\frac{1}{N}\sum_{i=1}^{N}X_i\right) = \frac{1}{N^2}\sum_{i=1}^{N}V(X_i) = \frac{1}{N}V(X) = \frac{1}{N}\Sigma$$

Now we check if $\hat{\mu}$ is unbiased estimate. In fact, we have:

$$E(\hat{\mu}) = E\left(\left(\hat{\Sigma}\Sigma_0^{-1} + NI\right)^{-1}\left(\hat{\Sigma}\Sigma_0^{-1}\mu_0 + N\bar{X}\right)\right) = \left(\hat{\Sigma}\Sigma_0^{-1} + NI\right)^{-1}\left(\hat{\Sigma}\Sigma_0^{-1}\mu_0 + NE(\bar{X})\right)$$

$$= \left(\hat{\Sigma}\Sigma_0^{-1} + NI\right)^{-1}\left(\hat{\Sigma}\Sigma_0^{-1}\mu_0 + \sum_{i=1}^{N}E(X_i)\right)$$

$$= \left(\hat{\Sigma}\Sigma_0^{-1} + NI\right)^{-1}\left(\hat{\Sigma}\Sigma_0^{-1}\mu_0 + NE(X)\right) = \left(\hat{\Sigma}\Sigma_0^{-1} + NI\right)^{-1}\left(\hat{\Sigma}\Sigma_0^{-1}\mu_0 + N\mu_0\right)$$

(Due to $E(X) = \mu_0$)

$$= \left(\hat{\Sigma}\Sigma_0^{-1} + NI\right)^{-1}\left(\hat{\Sigma}\Sigma_0^{-1} + NI\right)\mu_0 = \mu_0$$

Therefore, the estimate $\hat{\mu}$ is biased because the variable μ is not always to equal μ_0.

Now we try to check again if $\hat{\mu}$ is unbiased estimate with Cramer-Rao lower bound. The second-order partial derivative of $\ell(\Theta)$ regarding μ is:

$$\frac{\partial^2\ell(\Theta)}{\partial\mu^2} = \frac{\partial}{\partial\mu}\left(\frac{\partial\ell(\Theta)}{\partial\mu}\right) = \frac{\partial}{\partial\mu}\left(-\mu^T(\Sigma_0^{-1} + N\Sigma^{-1}) + \mu_0^T\Sigma_0^{-1} + \left(\sum_{i=1}^{N}X_i^T\right)\Sigma^{-1}\right)$$

$$= -(\Sigma_0^{-1} + N\Sigma^{-1})^T = -(\Sigma_0^{-1} + N\Sigma^{-1})$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

(Because Σ and Σ_0 are symmetric)

Cramer-Rao lower bound of $\hat{\mu}$ is:

$$CR(\hat{\mu}) = -\frac{1}{N}\left(\int_X \int_\Theta \frac{\partial^2 \ell(\Theta)}{\partial \mu^2} f(X,\Theta|\xi)\mathrm{d}X\mathrm{d}\Theta\right)^{-1}$$

$$= \frac{1}{N}\left(\int_X \int_\Theta (\Sigma_0^{-1} + N\Sigma^{-1}) f(X,\Theta|\xi)\mathrm{d}X\mathrm{d}\Theta\right)^{-1}$$

$$= \frac{1}{N}\left(\int_\Theta (\Sigma_0^{-1} + N\Sigma^{-1}) f(\Theta|\xi)\mathrm{d}\Theta\right)^{-1} = \frac{1}{N}\left(\int_\mu (\Sigma_0^{-1} + N\Sigma^{-1}) f(\mu|\mu_0,\Sigma_0)\mathrm{d}\Theta\right)^{-1}$$

$$= \frac{1}{N}(\Sigma_0^{-1} + N\Sigma^{-1})^{-1}$$

Variance of $\hat{\mu}$ is:

$$V(\hat{\mu}) = V\left((\Sigma\Sigma_0^{-1} + NI)^{-1}(\Sigma\Sigma_0^{-1}\mu_0 + N\bar{X})\right)$$

$$= V((\Sigma\Sigma_0^{-1} + NI)^{-1}\Sigma\Sigma_0^{-1}\mu_0 + N(\Sigma\Sigma_0^{-1} + NI)^{-1}\bar{X}) = V(N(\Sigma\Sigma_0^{-1} + NI)^{-1}\bar{X})$$

$$= N^2 V((\Sigma\Sigma_0^{-1} + NI)^{-1}\bar{X})$$

Because it is difficult to calculate $V(\hat{\mu})$, suppose we fix Σ so that $\hat{\Sigma} = \Sigma_0 = \Sigma$, we have:

$$V(\hat{\mu}) = N^2 V((\Sigma\Sigma_0^{-1} + NI)^{-1}\bar{X}) = N^2 V((\Sigma\Sigma^{-1} + NI)^{-1}\bar{X}) = N^2 V((I + NI)^{-1}\bar{X})$$

$$= N^2 V\left(\frac{1}{N+1}\bar{X}\right) = \frac{N^2}{(N+1)^2} V(\bar{X}) = \frac{N}{(N+1)^2}\Sigma$$

(Due to $V(\bar{X}) = \frac{1}{N}\Sigma$)

The Cramer-Rao lower bound of $\hat{\mu}$ is re-written as follows:

$$CR(\hat{\mu}) = \frac{1}{N}(\Sigma_0^{-1} + N\Sigma^{-1})^{-1} = \frac{1}{N}(\Sigma^{-1} + N\Sigma^{-1})^{-1} = \frac{1}{N}\left(\Sigma^{-1}(1+N)\right)^{-1} = \frac{1}{N(N+1)}\Sigma$$

Obviously, $\hat{\mu}$ is biased estimate due to $V(\hat{\mu}) \neq CR(\hat{\mu})$. In general, the estimate $\widehat{\Theta}$ in MAP is affected by the prior PDF $f(\Theta|\xi)$. Even though it is biased, it can be better than the one resulted from MLE because of valuable information in $f(\Theta|\xi)$. For instance, if fixing Σ, the variance of $\hat{\mu}$ from MAP $\left(\frac{N}{(N+1)^2}\Sigma\right)$ is "smaller" (lower bounded) than the one from MLE $\left(\frac{1}{N}\Sigma\right)$ ■

Now we skim through an introduction of EM algorithm. Suppose there are two spaces $\boldsymbol{X}$ 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 φ: $\boldsymbol{X} \to \boldsymbol{Y}$ and we denote $\varphi^{-1}(Y) = \{X \in \boldsymbol{X}: \varphi(X) = Y\}$ as all $X \in \boldsymbol{X}$ such that $\varphi(X) = Y$. We also denote $\boldsymbol{X}(Y) = \varphi^{-1}(Y)$. Let $f(X \mid \Theta)$ be the PDF of random variable $X \in \boldsymbol{X}$ and let $g(Y \mid \Theta)$ be the PDF of random variable $Y \in \boldsymbol{Y}$. Note, Y is also called observation. Equation 1.34 specifies $g(Y \mid \Theta)$ as integral of $f(X \mid \Theta)$ over $\varphi^{-1}(Y)$.

$$g(Y|\Theta) = \int_{\varphi^{-1}(Y)} f(X|\Theta)\mathrm{d}X \tag{1.34}$$

Where Θ is probabilistic parameter represented as a column vector, $\Theta = (\theta_1, \theta_2, \ldots, \theta_r)^T$ in which each θ_i is a particular parameter. According to viewpoint of Bayesian statistics, Θ is also random variable. As a convention, let Ω be the domain of Θ such that $\Theta \in \Omega$ and the dimension of Ω is r. For example, normal distribution has two particular parameters such as mean μ and variance σ^2 and so we have $\Theta = (\mu, \sigma^2)^T$. Note that, Θ can degrades into a scalar as $\Theta = \theta$. The conditional PDF of X given Y, denoted $k(X \mid Y, \Theta)$, is specified by equation 1.35.

$$k(X|Y,\Theta) = \frac{f(X|\Theta)}{g(Y|\Theta)} \tag{1.35}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

According to DLR (Dempster, Laird, & Rubin, 1977, p. 1), $\boldsymbol{X}$ is called *complete data* and the term "incomplete data" implies existence of $\boldsymbol{X}$ and $\boldsymbol{Y}$ where $\boldsymbol{X}$ is not observed directly and $\boldsymbol{X}$ is only known by the many-one mapping φ: $\boldsymbol{X} \rightarrow \boldsymbol{Y}$. In general, we only know $\boldsymbol{Y}$, $f(X \mid \Theta)$, and $k(X \mid Y, \Theta)$ and so our purpose is to estimate Θ based on such $\boldsymbol{Y}$, $f(X \mid \Theta)$, and $k(X \mid Y, \Theta)$. Like MLE approach, EM algorithm also maximizes the likelihood function to estimate Θ but the likelihood function in EM concerns $\boldsymbol{Y}$ and there are also some different aspects in EM which will be described later. Pioneers in EM algorithm firstly assumed that $f(X \mid \Theta)$ belongs to exponential family with note that many popular distributions such as normal, multinomial, and Poisson belong to exponential family (please see table 1.1). 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 PDF (s) have the same exponential form according to equation 1.36 (Dempster, Laird, & Rubin, 1977, p. 3):

$$f(X|\Theta) = b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta) \qquad (1.36)$$

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. For example, the sufficient statistic of normal distribution is $\tau(X) = (X, XX^T)^T$. Equation 1.36 expresses the canonical form of exponential family. Recall that Ω is the domain of Θ such that $\Theta \in \Omega$. Suppose that Ω is a convex set. If Θ is restricted only to Ω then, $f(X \mid \Theta)$ specifies a *regular exponential family*. If Θ lies in a curved sub-manifold Ω_0 of Ω then, $f(X \mid \Theta)$ specifies a *curved exponential family*. The $a(\Theta)$ is *partition function* for variable X, which is used for normalization.

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

As usual, a PDF is known as a popular form but its exponential family form (canonical form of exponential family) specified by equation 1.36 looks unlike popular form although they are the same. Therefore, parameter in popular form is different from parameter in exponential family form.

For example, multivariate normal distribution with theoretical mean μ and covariance matrix Σ of random variable $X = (x_1, x_2, \ldots, x_n)^T$ has PDF in popular form is:

$$f(X|\mu,\Sigma) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Hence, parameter in popular form is $\Theta = (\mu, \Sigma)^T$. Exponential family form of such PDF is:

$$f(X|\theta_1,\theta_2) = (2\pi)^{-\frac{n}{2}} * \exp\left((\theta_1,\theta_2)\begin{pmatrix}X\\XX^T\end{pmatrix}\right)\Big/\exp\left(-\frac{1}{4}\theta_1^T\theta_2^{-1}\theta_1 - \frac{1}{2}\log|-2\theta_2|\right)$$

Where,

$$\Theta = \begin{pmatrix}\theta_1\\\theta_2\end{pmatrix}$$

$$\theta_1 = \Sigma^{-1}\mu$$

$$\theta_2 = -\frac{1}{2}\Sigma^{-1}$$

$$b(X) = (2\pi)^{-\frac{n}{2}}$$

$$\tau(X) = \begin{pmatrix}X\\XX^T\end{pmatrix}$$

$$a(\Theta) = \exp\left(-\frac{1}{4}\theta_1^T\theta_2^{-1}\theta_1 - \frac{1}{2}\log|-2\theta_2|\right)$$

Hence, parameter in exponential family form is $\Theta = (\theta_1, \theta_2)^T$. Although, $f(X \mid \theta_1, \theta_2)$ looks unlike $f(X \mid \mu, \Sigma)$ but they are the same, $f(X \mid \theta_1, \theta_2) = f(X \mid \mu, \Sigma)$. In fact, we have:

$$\Theta^T\tau(X) = (\theta_1,\theta_2)\begin{pmatrix}X\\XX^T\end{pmatrix} = \left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right)\begin{pmatrix}X\\XX^T\end{pmatrix} = \mu^T\Sigma^{-1}X - \frac{1}{2}X^T\Sigma^{-1}X$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

We also have:

$$a(\Theta) = \exp\left(-\frac{1}{4}\theta_1^T\theta_2^{-1}\theta_1 - \frac{1}{2}\log|-2\theta_2|\right) = \exp\left(\frac{1}{2}\mu^T\Sigma^{-1}\Sigma\Sigma^{-1}\mu - \frac{1}{2}\log|\Sigma^{-1}|\right)$$

$$= \exp\left(\frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\log|\Sigma|\right) = |\Sigma|^{\frac{1}{2}} * \exp\left(\frac{1}{2}\mu^T\Sigma^{-1}\mu\right)$$

(Due to $|\Sigma^{-1}| = |\Sigma|^{-1}$)

Therefore,

$$f(X|\theta_1,\theta_2) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} * \exp\left(\mu^T\Sigma^{-1}X - \frac{1}{2}X^T\Sigma^{-1}X - \frac{1}{2}\mu^T\Sigma^{-1}\mu\right)$$

$$= (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}(X^T\Sigma^{-1}X - \mu^T\Sigma^{-1}X - \mu^T\Sigma^{-1}X + \mu^T\Sigma^{-1}\mu)\right)$$

$$= (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}(X^T\Sigma^{-1}X - \mu^T\Sigma^{-1}X - X^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu)\right)$$

(Because Σ is symmetric, $\mu^T\Sigma^{-1}X = X^T\Sigma^{-1}\mu$)

$$= (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}\left((X^T - \mu^T)\Sigma^{-1}X - (X^T - \mu^T)\Sigma^{-1}\mu\right)\right)$$

$$= (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}\left((X^T - \mu^T)\Sigma^{-1}(X - \mu)\right)\right)$$

$$= (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}(X - \mu)^T\Sigma^{-1}(X - \mu)\right) = f(X|\mu,\Sigma)\blacksquare$$

The exponential family form is used to represents all distributions belonging to exponential family as canonical form. Parameter in exponential family form is called exponential family parameter. As a convention, parameter Θ mentioned in EM algorithm is exponential family parameter if PDF belongs to exponential family and there is no additional information.

Table 1.1 shows some popular distributions belonging to exponential family along with their canonical forms (Wikipedia, Exponential family, 2016). In case of multivariate distributions, dimension of random variable $X = (x_1, x_2, \ldots, x_n)^T$ is n.

Distribution	Popular PDF	Exponential family parameter Θ	$\tau(X)$	$b(X)$	$a(\Theta)$
Multivariate normal	$f(X\|\mu,\Sigma) = \|2\pi\Sigma\|^{-\frac{1}{2}} * e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)}$	$\begin{pmatrix}\theta_1 = \Sigma^{-1}\mu \\ \theta_2 = -\frac{1}{2}\Sigma^{-1}\end{pmatrix}$	$\begin{pmatrix}X \\ XX^T\end{pmatrix}$	$(2\pi)^{-\frac{n}{2}}$	$\exp\left(-\frac{1}{4}\theta_1^T\theta_2^{-1}\theta_1 - \frac{1}{2}\log\|-2\theta_2\|\right)$
Multinomial	$f(X\|p_1,p_2,\ldots,p_n) = \frac{K!}{\prod_{j=1}^n(x_j!)}\prod_{k=1}^n p_j^{x_j}$ Where, $\sum_{j=1}^n p_j = 1$, $\sum_{j=1}^n x_j = K$, and $x_j \in \{0,1,\ldots,K\}$.	$\begin{pmatrix}\theta_1 = \log(p_1) \\ \theta_2 = \log(p_2) \\ \vdots \\ \theta_n = \log(p_n)\end{pmatrix}$	$\begin{pmatrix}x_1 \\ x_2 \\ \vdots \\ x_n\end{pmatrix}$	$\frac{K!}{\prod_{j=1}^n(x_j!)}$	1

Table 1.1. Some popular distributions belonging to exponential family

It is necessary to survey some features of exponential family. The first-order derivative of $\log(a(\Theta))$ is expectation of $\tau(X)$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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,

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

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

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

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.

$$a(\Theta|Y) = \int_{\varphi^{-1}(Y)} b(X)\exp\big(\Theta^T\tau(X)\big)\mathrm{d}X$$

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

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

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

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

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

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

Note that $k(X \mid Y, \Theta)$ is determined on $X \in \varphi^{-1}(Y)$. Of course, we have:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

$$\log'\left(a(\Theta|Y)\right) = E(\tau(X)|Y,\Theta) = \int_{\varphi^{-1}(Y)} \tau(X)k(X|Y,\Theta)\mathrm{d}X$$

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

$$\log''\left(a(\Theta|Y)\right) = V(\tau(X)|Y,\Theta)$$
$$= \int_{\varphi^{-1}(Y)} \left(\tau(X) - E(\tau(X)|Y,\Theta)\right)\left(\tau(X) - E(\tau(X)|Y,\Theta)\right)^T k(X|Y,\Theta)\mathrm{d}X$$

Where $V(\tau(X) \mid Y, \Theta)$ is central covariance matrix of $\tau(X)$ given observed Y. Table 1.2 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.

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

Table 1.2. 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 related to 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).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

Example 1.1. Rao (Rao, 1955) presents observed 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 PDF 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:

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

Observed 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 PDF 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 φ between X and Y is $y_1 = \varphi(x_1, x_2) = x_1 + x_2$. Therefore $g(Y \mid \theta)$ is sum of $f(X \mid \theta)$ over x_1 and x_2 such that $x_1 + x_2 = y_1$ according to equation 1.34. In other words, $g(Y \mid \theta)$ is resulted from summing $f(X \mid \theta)$ over all (x_1, x_2) pairs such as $(0, 125)$, $(1, 124)$,…, $(125, 0)$ and then substituting $(18, 20, 34)$ for (x_3, x_4, x_5) because of $y_1 = 125$ from observed Y.

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

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

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

Given $p_{y1} = 1/2 + \theta/4$, which implies that:

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

When $y_1 = 125$, we have:

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

This suggests us to select:

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

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

Please pay attention that the expectation $y_1^{(t)} = E(y_1 \mid Y, \theta^{(t)})$ gets value 125 when y_1 is evaluated as $y_1 = 125$ and the probability corresponding to y_1 gets maximal as $1/2 + \theta^{(t)}/4 = 1$.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

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

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

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

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

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

So we have:

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

Where,

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

For example, given the initial $\theta^{(1)} = 0.5$, at the first iteration, we have:

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

$$\theta^{(2)} = \frac{x_2^{(1)} + 34}{x_2^{(1)} + 72} = \frac{25 + 34}{25 + 72} = 0.6082$$

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

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

Table 1.3 (Dempster, Laird, & Rubin, 1977, p. 3) lists estimates of θ over five iterations (t =1, 2, 3, 4, 5) with note that $\theta^{(1)}$ is initialized arbitrarily and $\theta^* = \theta^{(5)} = \theta^{(6)}$ is determined at the 5th iteration. The third column gives deviation θ^* and $\theta^{(t)}$ whereas the fourth column gives the ratio of successive deviations. Later on, we will know that such ratio implies convergence rate.

t	$\theta^{(t)}$	$\theta^* - \theta^{(t)}$	$(\theta^* - \theta^{(t+1)})$ / $(\theta^* - \theta^{(t)})$
1	$\theta^{(1)} = 0.5$	0.1268	0.1465
	$\theta^{(2)} = 0.6082$	0.0186	0.1346
2	$\theta^{(2)} = 0.6082$	0.0186	0.1346
	$\theta^{(3)} = 0.6243$	0.0025	0.1330
3	$\theta^{(3)} = 0.6243$	0.0025	0.1330
	$\theta^{(4)} = 0.6265$	0.0003	0.1328
4	$\theta^{(4)} = 0.6265$	0.0003	0.1328
	$\theta^{(5)} = 0.6268$	0	0.1328
5	$\theta^{(5)} = 0.6268$	0	0.1328
	$\theta^{(6)} = 0.6268$	0	0.1328

Table 1.3. EM algorithm in simple case

For example, at the first iteration, we have:

$$\theta^* - \theta^{(1)} = 0.6268 - 0.5 = 0.1268$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\frac{\theta^* - \theta^{(2)}}{\theta^* - \theta^{(1)}} = \frac{\theta^{(2)} - \theta^*}{\theta^{(1)} - \theta^*} = \frac{0.6082 - 0.6268}{0.5 - 0.6268} = 0.1465$$

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 PDF $f(X \mid \Theta)$ 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 $\Theta^{(t)}$, with note that $f(X \mid \Theta)$ belongs to regular exponential family.

E-step:

We calculate current value $\tau^{(t)}$ of the sufficient statistic $\tau(X)$ from observed Y and current parameter $\Theta^{(t)}$ according to equation 2.6:

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

M-step:

Basing on $\tau^{(t)}$, we determine the next parameter $\Theta^{(t+1)}$ as solution of equation 2.3:

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

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 given regular exponential PDF $f(X|\Theta)$

EM algorithm stops if two successive estimates are equal, $\Theta^* = \Theta^{(t)} = \Theta^{(t+1)}$, at some t^{th} iteration. At that time we conclude that Θ^* is the optimal estimate of EM process. Please see table 1.2 to know how to calculate $E(\tau(X) \mid \Theta^{(t)})$ and $E(\tau(X) \mid Y, \Theta^{(t)})$. As a convention, the estimate of parameter Θ resulted from EM process is denoted Θ^* instead of $\widehat{\Theta}$ in order to emphasize that Θ^* is solution of optimization problem.

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

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

Where,

$$L(\Theta) = \log\left(g(Y|\Theta)\right)$$

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

$$l(\Theta) = \log\left(f(X|\Theta)\right) = \log\left(b(X)\right) + \Theta^T\tau(X) - \log\left(a(\Theta)\right) \quad (2.1)$$

By referring to table 1.2, the first-order derivative of $l(\Theta)$ is:

$$\frac{\mathrm{d}l(\Theta)}{\mathrm{d}\Theta} = \frac{\mathrm{d}\log\left(f(Y|\Theta)\right)}{\mathrm{d}\Theta} = \tau(X) - \log'\left(a(\Theta)\right) = \tau(X) - E(\tau(X)|\Theta) \quad (2.2)$$

We set the first-order derivative of $l(\Theta)$ to be zero with expectation that $l(\Theta)$ will be maximized. Therefore, the optimal estimate Θ^* 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 Θ but $\tau(X)$ is still dependent on X. Let $\tau^{(t)}$ be value of $\tau(X)$ at the t^{th} iteration of EM process, candidate for the best estimate of Θ is solution of equation 2.3 according to M-step.

$$E(\tau(X)|\Theta) = \tau^{(t)} \quad (2.3)$$

Where,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$E(\tau(X)|\Theta) = \int_{X} f(X|\Theta)\tau(X)\mathrm{d}X$$

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

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

Where,

$$L(\Theta) = \log\big(g(Y|\Theta)\big) = \log\left(\int_{\varphi^{-1}(Y)} f(X|\Theta)\mathrm{d}X\right) \quad (2.4)$$

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 1.2, the first-order derivative of $L(\Theta)$ is:

$$\frac{\mathrm{d}L(\Theta)}{\mathrm{d}\Theta} = -\log'\big(a(\Theta)\big) + \log'\big(a(\Theta|Y)\big) = -E(\tau(X)|\Theta) + E(\tau(X)|Y,\Theta) \quad (2.5)$$

We set the first-order derivative of $L(\Theta)$ to be zero with expectation that $L(\Theta)$ will be maximized, 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^{th} iteration of EM process. Derived from the equality above, the value $\tau^{(t)}$ is calculated as seen in equation 2.6.

$$\tau^{(t)} = E\big(\tau(X)\big|Y,\Theta^{(t)}\big) \quad (2.6)$$

Where,

$$E\big(\tau(X)\big|Y,\Theta^{(t)}\big) = \int_{\varphi^{-1}(Y)} k\big(X\big|Y,\Theta^{(t)}\big)\tau(X)\mathrm{d}X$$

Equation 2.6 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 2.3 (Dempster, Laird, & Rubin, 1977, p. 5). This means that Θ^* is the optimal estimate of EM process because Θ^* is solution of the equation:

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

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

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

The EM algorithm shown in table 2.1 is totally exact with assumption that $f(X|\Theta)$ belongs to regular exponential family. If $f(X|\Theta)$ is not regular, the maximal point (maximizer) of the log-likelihood function $l(\Theta)$ is not always the stationary point Θ^* so that the first-order derivative of $l(\Theta)$ is zero, $l'(\Theta^*) = 0$. However, if $f(X|\Theta)$ belongs to curved exponential family, the M-step of the EM algorithm shown in table 2.1 is modified as follows (Dempster, Laird, & Rubin, 1977, p. 5):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\Theta^{(t+1)} = \underset{\Theta\in\Omega_0}{\operatorname{argmax}}\, l(\Theta) = \underset{\Theta\in\Omega_0}{\operatorname{argmax}}\, l\left(\Theta\middle|\tau^{(t)}\right) = \underset{\Theta\in\Omega_0}{\operatorname{argmax}}\left(\Theta^T\tau^{(t)} - \log\left(a(\Theta)\right)\right) \quad (2.7)$$

Where $\tau^{(t)}$ is calculated by equation 2.6 in E-step. This means that, in more general manner, the maximizer $\Theta^{(t+1)}$ will be found by some way. Recall that if Θ lies in a curved sub-manifold Ω_0 of Ω where Ω is the domain of Θ then, $f(X \mid \Theta)$ belongs to curved exponential family.

In general, given exponential family, within simple EM algorithm, E-step aims to calculate the current sufficient statistic $\tau^{(t)}$ that the log-likelihood function $L(\Theta^{(t)})$ gets maximal with such $\tau^{(t)}$ at current $\Theta^{(t)}$ given Y whereas M-step aims to maximize the log-likelihood function $l(\Theta)$ given $\tau^{(t)}$, as seem in table 2.2. Note, in table 2.2, $f(X|\Theta)$ belongs to curved exponential family but it is not necessary to be regular.

E-step:

Given observed Y and current $\Theta^{(t)}$, current value $\tau^{(t)}$ of the sufficient statistic $\tau(X)$ is the value that the log-likelihood function $L(\Theta^{(t)})$ gets maximal with such $\tau^{(t)}$. Concretely, suppose Θ^* be a maximizer of $L(\Theta)$ given Y where $L(\Theta)$ is specified by equation 2.4.

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

Suppose Θ^* is formulated as function of $\tau(X)$, for instance, $\Theta^* = h(\tau(X))$ with note that Θ^* is not evaluated because $\tau(X)$ is not evaluated. Thus, the equation $\Theta^* = h(\tau(X))$ is only symbolic formula. Let $\tau^{(t)}$ be a value of $\tau(X)$ such that $\Theta^{(t)} = h(\tau(X))$. This means $\tau^{(t)} \in \left\{\tau(X): \Theta^{(t)} = h\left(\tau(X)\right)\right\}$ with note that Θ^* is replaced by $\Theta^{(t)}$. If $h(\tau(X))$ is invertible, $\tau^{(t)} = h^{-1}(\Theta^{(t)})$.

If the PDF $f(X|\Theta)$ belongs to regular exponential family, $\tau^{(t)}$ is calculated more easily according to equation 2.6, given Y and $\Theta^{(t)}$.

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

Where,

$$E\left(\tau(X)\middle|Y, \Theta^{(t)}\right) = \int_{\varphi^{-1}(Y)} k\left(X\middle|Y, \Theta^{(t)}\right)\tau(X)\mathrm{d}X$$

M-step:

Basing on $\tau^{(t)}$, we determine the next parameter $\Theta^{(t+1)}$ by maximizing the log-likelihood function $l(\Theta)$ given $\tau^{(t)}$, where $l(\Theta)$ is specified by equation 2.1. Actually, the sufficient statistic $\tau^{(t)}$ calculated in E-step is substituted for unobserved $\tau(X)$ in $l(\Theta)$ so that it is possible to maximize $l(\Theta)$ with subject to Θ.

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}}\, l\left(\Theta\middle|\tau^{(t)}\right)$$

If the PDF $f(X|\Theta)$ belongs to regular exponential family, $\Theta^{(t+1)}$ is solution of equation 2.3 given $\tau^{(t)}$.

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

Where,

$$E(\tau(X)|\Theta) = \int_{\boldsymbol{X}} f(X|\Theta)\tau(X)\mathrm{d}X$$

If the PDF $f(X|\Theta)$ belongs to curved exponential family, $\Theta^{(t+1)}$ is determined by equation 2.7 given $\tau^{(t)}$.

$$\Theta^{(t+1)} = \underset{\Theta\in\Omega_0}{\operatorname{argmax}}\left(\Theta^T\tau^{(t)} - \log\left(a(\Theta)\right)\right)$$

Table 2.2. E-step and M-step of EM algorithm given exponential PDF $f(X|\Theta)$

EM algorithm stops if two successive estimates are equal, $\Theta^* = \Theta^{(t)} = \Theta^{(t+1)}$, at some t^{th} iteration. At that time, Θ^* is the optimal estimate of EM process, which is an optimizer of $L(\Theta)$.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

Going back example 1.1, 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 $\theta^{(t)}$ in E-step according to equation 2.6.

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

Given $p_{y1} = 1/2 + \theta/4$, which implies that:

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

Because the probability of y_1 is $1/2 + \theta/4$ and y_1 is sum of x_1 and x_2, let $p_{x_1|y_1}$ be conditional probability of x_1 given y_1 and let $p_{x_2|y_1}$ be conditional probability of x_2 given y_1 such that

$$p_{x_1|y_1} = \frac{P(x_1, y_1)}{p_{y_1}} = \frac{P(x_1, y_1)}{1/2 + \theta/4}$$

$$p_{x_2|y_1} = \frac{P(x_2, y_1)}{p_{y_1}} = \frac{P(x_2, y_1)}{1/2 + \theta/4}$$

$$p_{x_1|y_1} + p_{x_2|y_1} = 1$$

Where $P(x_1, y_1)$ and $P(x_2, y_1)$ are joint probabilities of (x_1, y_1) and (x_2, y_1), respectively. We can select $P(x_1, y_1) = 1/2$ and $P(x_2, y_1) = \theta/4$, which implies:

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

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

Such that

$$x_1^{(t)} + x_2^{(t)} = y_1^{(t)}$$

Note, we can select alternately as $P(x_1, y_1) = P(x_2, y_1) = (1/2 + \theta/4) / 2$, for example but fixing $P(x_1, y_1)$ as 1/2 is better because the next estimate $\theta^{(t+1)}$ known later depends only on $x_2^{(t)}$.

When y_1 is evaluated as $y_1 = 125$, we obtain:

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

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

Essentially, equation 2.3 specifying M-step is result of maximizing the log-likelihood function $l(\Theta)$. This log-likelihood function is:

$$l(\Theta) = \log\left(f(X|\theta)\right)$$

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

$$+ (x_2 + x_5)\log(\theta) + (x_3 + x_4)\log(1 - \theta)$$

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

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

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: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

Where,

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

For example, given the initial $\theta^{(1)} = 0.5$, at the first iteration, we have:

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

$$\theta^{(2)} = \frac{x_2^{(1)} + 34}{x_2^{(1)} + 72} = \frac{25 + 34}{25 + 72} = 0.6082$$

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

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

Table 1.3 (Dempster, Laird, & Rubin, 1977, p. 3) show resulted estimation ■

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 2.8 (Dempster, Laird, & Rubin, 1977, p. 6).

$$Q(\Theta'|\Theta) = E\left(\log\left(f(X|\Theta')\right)\middle|Y,\Theta\right) = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\left(f(X|\Theta')\right)\mathrm{d}X \qquad (2.8)$$

The two steps of generalized EM (*GEM*) algorithm aim to maximize $Q(\Theta \mid \Theta^{(t)})$ at some t^{th} iteration as seen in table 2.3 (Dempster, Laird, & Rubin, 1977, p. 6).

E-step: The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current parameter $\Theta^{(t)}$, according to equation 2.8. Actually, $Q(\Theta \mid \Theta^{(t)})$ is formulated as function of Θ. *M-step*: The next parameter $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with subject to Θ. Note that $\Theta^{(t+1)}$ will become current parameter at the next iteration (the $(t+1)^{\text{th}}$ iteration).

Table 2.3. E-step and M-step of GEM algorithm

DLR proved that GEM algorithm converges at some t^{th} iteration. At that time, $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ is the optimal estimate of EM process, which is an optimizer of $L(\Theta)$.

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

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 the optimization problem which is dependent on applications. For example, the estimate $\Theta^{(t+1)}$ can be solution of the equation created by setting the first-order derivative of $Q(\Theta \mid \Theta^{(t)})$ regarding Θ to be zero. If solving such equation is too complex, some popular methods to solve optimization problem are Newton-Raphson (Burden & Faires, 2011, pp. 67-71), gradient descent (Ta, 2014), and Lagrange duality (Wikipedia, Karush–Kuhn–Tucker conditions, 2014).

GEM algorithm still aims to maximize the log-likelihood function $L(\Theta)$ specified by equation 2.4, which is explained here. Following is proof of equation 2.8. Suppose the current parameter is Θ after some iteration. Next we must find out the new estimate Θ^* that maximizes the next log-likelihood function $L(\Theta')$.

$$\Theta^* = \underset{\Theta'}{\operatorname{argmax}}\, L(\Theta') = \underset{\Theta'}{\operatorname{argmax}} \log\left(g(Y|\Theta')\right)$$

The next log-likelihood function $L(\Theta')$ is re-written as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$L(\Theta') = \log\left(\int_{\varphi^{-1}(Y)} f(X|\Theta')\mathrm{d}X\right) = \log\left(\int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\frac{f(X|\Theta')}{k(X|Y,\Theta)}\mathrm{d}X\right)$$

Due to

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

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

$$\log\left(\int_x u(x)v(x)\mathrm{d}x\right) \geq \int_x u(x)\log\big(v(x)\big)\mathrm{d}x$$

$$\text{where} \int_x u(x)\mathrm{d}x = 1$$

into $L(\Theta')$, we have (Sean, 2009, p. 6):

$$L(\Theta') \geq \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\left(\frac{f(X|\Theta')}{k(X|Y,\Theta)}\right)\mathrm{d}X$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\left(\log\big(f(X|\Theta')\big) - \log\big(k(X|Y,\Theta)\big)\right)\mathrm{d}X$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X - \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta)\big)\mathrm{d}X$$

$$= Q(\Theta'|\Theta) - H(\Theta|\Theta)$$

Where,

$$Q(\Theta'|\Theta) = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X$$

$$H(\Theta'|\Theta) = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta')\big)\mathrm{d}X$$

The lower-bound of $L(\Theta')$ is defined as follows:

$$lb(\Theta' \mid \Theta) = Q(\Theta' \mid \Theta) - H(\Theta \mid \Theta)$$

Of course, we have:

$$L(\Theta') \geq lb(\Theta' \mid \Theta)$$

Suppose at some t^{th} iteration, when the current parameter is $\Theta^{(t)}$, the lower-bound of $L(\Theta)$ is re-written:

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

Of course, we have:

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

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

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

Indeed, we have:

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

$$= \int_{\varphi^{-1}(Y)} k\big(X\big|Y,\Theta^{(t)}\big)\log\left(f\big(X\big|\Theta^{(t)}\big)\right)\mathrm{d}X - \int_{\varphi^{-1}(Y)} k\big(X\big|Y,\Theta^{(t)}\big)\log\left(k\big(X\big|Y,\Theta^{(t)}\big)\right)\mathrm{d}X$$

$$= \int_{\varphi^{-1}(Y)} k\big(X\big|Y,\Theta^{(t)}\big)\log\left(\frac{f\big(X\big|\Theta^{(t)}\big)}{k\big(X\big|Y,\Theta^{(t)}\big)}\right)\mathrm{d}X$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$= \int_{\varphi^{-1}(Y)} k\left(X|Y,\Theta^{(t)}\right)\log\left(g\left(Y|\Theta^{(t)}\right)\right)\mathrm{d}X = \log\left(g\left(Y|\Theta^{(t)}\right)\right)\int_{\varphi^{-1}(Y)} k\left(X|Y,\Theta^{(t)}\right)\mathrm{d}X$$
$$= \log\left(g\left(Y|\Theta^{(t)}\right)\right) = L\left(\Theta^{(t)}\right)$$

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

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}}\, lb\left(\Theta|\Theta^{(t)}\right) = \underset{\Theta}{\operatorname{argmax}}\, Q\left(\Theta|\Theta^{(t)}\right)$$

Note, $H(\Theta^{(t)} \mid \Theta^{(t)})$ is constant with regard to Θ. The lower bound is increased after every iteration. As a result, the maximizer Θ^* of the final lower-bound after many iterations will be expected as a maximizer of $L(\Theta)$ in final.

Therefore, the two steps of GEM is interpreted with regard to the lower bound $lb(\Theta \mid \Theta^{(t)})$ as seen in table 2.4.

E-step: The lower bound $lb(\Theta \mid \Theta^{(t)})$ is re-calculated based on $Q(\Theta \mid \Theta^{(t)})$. *M-step*: The next parameter $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ which is also a maximizer of $lb(\Theta \mid \Theta^{(t)})$ because $H(\Theta^{(t)} \mid \Theta^{(t)})$ is constant. $\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}}\, lb\left(\Theta|\Theta^{(t)}\right) = \underset{\Theta}{\operatorname{argmax}}\, Q\left(\Theta|\Theta^{(t)}\right)$ 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 2.4. 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. Figure 2.1 (Borman, 2004, p. 7) shows relationship between the log-likelihood function $L(\Theta)$ and its lower-bound $lb(\Theta \mid \Theta^{(t)})$.

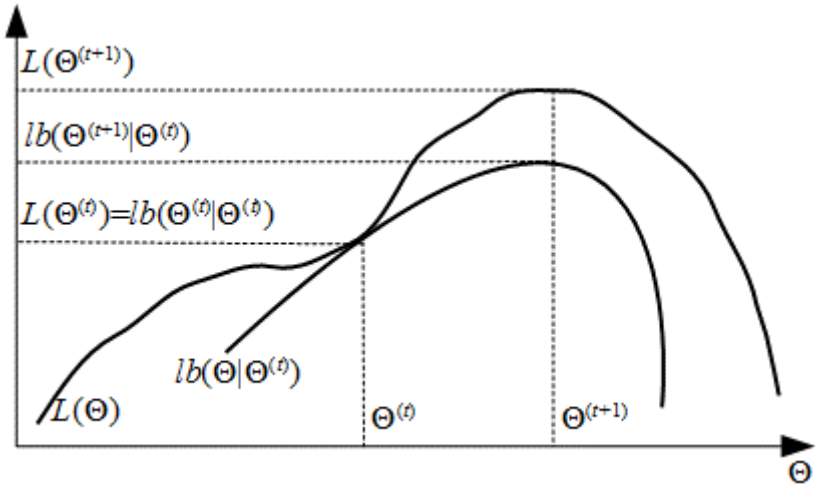

Figure 2.1. Relationship between the log-likelihood function and its lower-bound

Now ideology of GEM is explained in detail ■

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

The next section focuses on 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 modified by equation 2.9.

$$Q(\Theta'|\Theta) = E\big(\log\big(b(X)\big)\big|Y,\Theta\big) + (\Theta')^T\tau_\Theta - \log\big(a(\Theta')\big) \tag{2.9}$$

Where,

$$E\big(\log\big(b(X)\big)\big|Y,\Theta\big) = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(b(X)\big)\mathrm{d}X$$

$$\tau_\Theta = E(\tau(X)|Y,\Theta) = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\tau(X)\mathrm{d}X$$

Following is a proof of equation 2.9.

$$Q(\Theta'|\Theta) = E\big(\log\big(f(X|\Theta')\big)\big|Y,\Theta\big) = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X$$

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

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\left(\log\big(b(X)\big) + (\Theta')^T\tau(X) - \log\big(a(\Theta')\big)\right)\mathrm{d}X$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(b(X)\big)\mathrm{d}X + \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)(\Theta')^T\tau(X)\mathrm{d}X - \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(a(\Theta')\big)\mathrm{d}X$$

$$= E\big(\log\big(b(X)\big)\big|Y,\Theta\big) + (\Theta')^T\int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\tau(X)\mathrm{d}X - \log\big(a(\Theta')\big)$$

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

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

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

If $f(X|\Theta)$ belongs to regular exponential family, $Q(\Theta' \mid \Theta)$ gets maximal at the stationary point Θ^* so that the first-order derivative of $Q(\Theta' \mid \Theta)$ is zero. By referring to table 1.2, the first-order derivative of $Q(\Theta' \mid \Theta)$ with regard to Θ' is:

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

Let $\tau^{(t)}$ be the value of τ_Θ at the t^{th} iteration.

$$\tau^{(t)} = E\big(\tau(X)\big|Y,\Theta^{(t)}\big) = \int_{\varphi^{-1}(Y)} k\big(X\big|Y,\Theta^{(t)}\big)\tau(X)\mathrm{d}X$$

The equation above is indeed equation 2.6. The next parameter $\Theta^{(t+1)}$ is determined at M-step as solution of the following equation.

$$\frac{\mathrm{d}Q(\Theta'|\Theta)}{\mathrm{d}\Theta'} = \tau^{(t)} - E(\tau(X)|\Theta) = 0$$

The equation above is indeed equation 2.3. If $f(X|\Theta)$ belongs to curved exponential family, $\Theta^{(t+1)}$ is determined as follows:

$$\Theta^{(t+1)} = \underset{\Theta'}{\operatorname{argmax}}\, Q(\Theta'|\Theta) = \underset{\Theta'}{\operatorname{argmax}}\left((\Theta')^T\tau^{(t)} - \log\big(a(\Theta')\big)\right)$$

The equation above is indeed equation 2.7. Therefore, GEM shown in table 2.3 degrades into EM shown in table 2.1 and table 2.2 if $f(X|\Theta)$ belongs to exponential family. Of course, this

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

recognition is trivial. Example 1.1 is also a good example for GEM when multinomial distribution belongs to exponential family and then we apply equation 2.7 into maximizing $Q(\Theta' \mid \Theta)$.

In practice, if Y is observed as particular N observations $Y_1, Y_2,\ldots, Y_N$. Let $\mathcal{Y} = \{Y_1, Y_2,\ldots, Y_N\}$ be the observed sample of size N with note that all Y_i (s) are mutually independent and identically distributed (iid). Given an observation Y_i, there is an associated random variable X_i. All X_i (s) are iid and they are not existent in fact. Each $X_i \in \boldsymbol{X}$ is a random variable like X. Of course, the domain of each X_i is $\boldsymbol{X}$. Let $\mathcal{X} = \{X_1, X_2,\ldots, X_N\}$ be the set of associated random variables. Because all X_i (s) are iid, the joint PDF of $\mathcal{X}$ is determined as follows:

$$f(\mathcal{X}|\Theta) = f(X_1, X_2, \ldots, X_N|\Theta) = \prod_{i=1}^{N} f(X_i|\Theta)$$

Because all X_i (s) are iid and each Y_i is associated with X_i, the conditional joint PDF of $\mathcal{X}$ given $\mathcal{Y}$ is determined as follows:

$$k(\mathcal{X}|\mathcal{Y},\Theta) = k(X_1, X_2, \ldots, X_N|Y_1, Y_2, \ldots, Y_N, \Theta) = \prod_{i=1}^{N} k(X_i|Y_1, Y_2, \ldots, Y_N, \Theta) = \prod_{i=1}^{N} k(X_i|Y_i, \Theta)$$

The conditional expectation $Q(\Theta' \mid \Theta)$ given samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ is determined as follows:

$$Q(\Theta'|\Theta) = \int_{\varphi^{-1}(\mathcal{Y})} k(\mathcal{X}|\mathcal{Y},\Theta)\log\big(f(\mathcal{X}|\Theta')\big)\mathrm{d}\mathcal{X}$$

$$= \int_{\varphi^{-1}(Y_1)} \int_{\varphi^{-1}(Y_2)} \ldots \int_{\varphi^{-1}(Y_N)} \left(\prod_{j=1}^{N} k\big(X_j|Y_j,\Theta\big)\right) * \left(\log\left(\prod_{i=1}^{N} f(X_i|\Theta')\right)\right) \mathrm{d}X_N \ldots \mathrm{d}X_2\, \mathrm{d}X_1$$

$$= \int_{\varphi^{-1}(Y_1)} \int_{\varphi^{-1}(Y_2)} \ldots \int_{\varphi^{-1}(Y_N)} \left(\prod_{j=1}^{N} k\big(X_j|Y_j,\Theta\big)\right) * \left(\sum_{i=1}^{N} \log\big(f(X_i|\Theta')\big)\right) \mathrm{d}X_N \ldots \mathrm{d}X_2\, \mathrm{d}X_1$$

$$= \int_{\varphi^{-1}(Y_1)} \int_{\varphi^{-1}(Y_2)} \ldots \int_{\varphi^{-1}(Y_N)} \sum_{i=1}^{N} \left(\prod_{j=1}^{N} k\big(X_j|Y_j,\Theta\big)\right) * \log\big(f(X_i|\Theta')\big)\, \mathrm{d}X_N \ldots \mathrm{d}X_2\, \mathrm{d}X_1$$

$$= \sum_{i=1}^{N} \int_{\varphi^{-1}(Y_1)} \int_{\varphi^{-1}(Y_2)} \ldots \int_{\varphi^{-1}(Y_N)} \log\big(f(X_i|\Theta')\big) * \prod_{j=1}^{N} k\big(X_j|Y_j,\Theta\big)\, \mathrm{d}X_N \ldots \mathrm{d}X_2\, \mathrm{d}X_1$$

(Suppose $f(X_i \mid \Theta)$ and $k(X_j \mid Y_j, \Theta)$ are analytic functions)

$$= \sum_{i=1}^{N} \int_{\varphi^{-1}(Y_1)} \int_{\varphi^{-1}(Y_2)} \ldots \int_{\varphi^{-1}(Y_N)} \int_{X} \delta(X, X_i)\log\big(f(X|\Theta')\big) * \prod_{j=1}^{N} k\big(X_j|Y_j,\Theta\big)\, \mathrm{d}X\, \mathrm{d}X_N \ldots \mathrm{d}X_2\, \mathrm{d}X_1$$

$$\left(\delta(X, X_i) = \begin{cases} 1 \text{ if } X = X_i \\ 0 \text{ if } X \neq X_i \end{cases} \Rightarrow \int_{X} \delta(X, X_i)u(X)\mathrm{d}X = u(X_i) \right.$$

$$\left. \begin{array}{c} \text{according to Riemann integral} \\ \text{with note that the domain of } X \text{ and } X_i \text{ is } \boldsymbol{X} \end{array} \right)$$

$$= \sum_{i=1}^{N} \int_{X} \int_{\varphi^{-1}(Y_1)} \int_{\varphi^{-1}(Y_2)} \ldots \int_{\varphi^{-1}(Y_N)} \delta(X, X_i)\log\big(f(X|\Theta')\big) * \prod_{j=1}^{N} k\big(X_j|Y_j,\Theta\big)\, \mathrm{d}X_N \ldots \mathrm{d}X_2\, \mathrm{d}X_1\, dX$$

$$= \sum_{i=1}^{N} \int_{X} \int_{\varphi^{-1}(Y_1),\varphi^{-1}(Y_2),\ldots,\varphi^{-1}(Y_N)} \delta(X, X_i)\log\big(f(X|\Theta')\big) * \prod_{j=1}^{N} k\big(X_j|Y_j,\Theta\big)\, \mathrm{d}X_N \mathrm{d}X_2 \mathrm{d}X_1 \ldots dX$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$= \sum_{i=1}^{N} \int_{X} \log\left(f(X|\Theta')\right) * \int_{\varphi^{-1}(Y_1),\varphi^{-1}(Y_2),\ldots,\varphi^{-1}(Y_N)} \delta(X, X_i) * \prod_{j=1}^{N} k\left(X_j|Y_j, \Theta\right) \mathrm{d}X_N \mathrm{d}X_2 \mathrm{d}X_1 \ldots dX$$

$$= \sum_{i=1}^{N} \int_{X} \log\left(f(X|\Theta')\right) * \int_{\varphi^{-1}(Y_1),\varphi^{-1}(Y_2),\ldots,\varphi^{-1}(Y_N)} \delta(X, X_i) k(X_i|Y_i, \Theta) * \prod_{j=1, j\neq i}^{N} k\left(X_j|Y_j, \Theta\right) \mathrm{d}X_N \mathrm{d}X_2 \mathrm{d}X_1 \ldots dX$$

$$= \sum_{i=1}^{N} \int_{X} \log\left(f(X|\Theta')\right) * \int_{\substack{\varphi^{-1}(Y_1),\varphi^{-1}(Y_2),\ldots,\varphi^{-1}(Y_{i-1}),\\ \varphi^{-1}(Y_i),\varphi^{-1}(Y_{i+1}),\ldots,\varphi^{-1}(Y_N)}} \delta(X, X_i) k(X_i|Y_i, \Theta) * \prod_{j=1, j\neq i}^{N} k\left(X_j|Y_j, \Theta\right) \mathrm{d}X_N \ldots \mathrm{d}X_{i+1} \mathrm{d}X_i \mathrm{d}X_{i-1} \ldots \mathrm{d}X_2 \mathrm{d}X_1 \ldots \mathrm{d}X$$

$$= \sum_{i=1}^{N} \int_{X} \log\left(f(X|\Theta')\right) * \int_{\varphi^{-1}(Y_1),\varphi^{-1}(Y_2),\ldots\varphi^{-1}(Y_{i-1})} \int_{\varphi^{-1}(Y_i)} \delta(X, X_i) k(X_i|Y_i, \Theta) * \int_{\varphi^{-1}(Y_{i+1}),\ldots,\varphi^{-1}(Y_N)} \prod_{j=1, j\neq i}^{N} k\left(X_j|Y_j, \Theta\right) \mathrm{d}X_N \ldots \mathrm{d}X_{i+1}\, \mathrm{d}X_i\, \mathrm{d}X_{i-1} \ldots \mathrm{d}X_2 \mathrm{d}X_1\, \mathrm{d}X$$

$$= \sum_{i=1}^{N} \int_{X} \log\left(f(X|\Theta')\right) * \left(\int_{\varphi^{-1}(Y_i)} \delta(X, X_i) k(X_i|Y_i, \Theta) \mathrm{d}X_i \right) * \int_{\substack{\varphi^{-1}(Y_1),\varphi^{-1}(Y_2),\ldots,\\ \varphi^{-1}(Y_{i-1}),\varphi^{-1}(Y_{i+1}),\ldots,\varphi^{-1}(Y_N)}} \prod_{j=1, j\neq i}^{N} k\left(X_j|Y_j, \Theta\right) \mathrm{d}X_N \ldots \mathrm{d}X_{i+1} \mathrm{d}X_{i-1} \ldots \mathrm{d}X_2 \mathrm{d}X_1\, \mathrm{d}X$$

$$= \sum_{i=1}^{N} \int_{X} \log\left(f(X|\Theta')\right) * \left(\int_{\varphi^{-1}(Y_i)} \delta(X, X_i) k(X_i|Y_i, \Theta) \mathrm{d}X_i \right) * \left(\prod_{j=1, j\neq i}^{N} \int_{\varphi^{-1}(Y_j)} k\left(X_j|Y_j, \Theta\right) \mathrm{d}X_j \right) \mathrm{d}X$$

$$= \sum_{i=1}^{N} \int_{X} \log\left(f(X|\Theta')\right) * \left(\int_{\varphi^{-1}(Y_i)} \delta(X, X_i) k(X_i|Y_i, \Theta) \mathrm{d}X_i \right) \mathrm{d}X$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\left(\text{Due to } \int_{\varphi^{-1}(Y_j)} k(X_j|Y_j,\Theta)\mathrm{d}X_j = 1\right)$$

$$= \sum_{i=1}^{N} \int_{\varphi^{-1}(Y_i)} \int_{X} \delta(X,X_i)k(X_i|Y_i,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X\,\mathrm{d}X_i$$

(Suppose $f(X_i \mid \Theta)$ and $k(X_j \mid Y_j, \Theta)$ are analytic functions)

By taking Riemann integral on $\int_X \delta(X,X_i)k(X_i|Y_i,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X$, we have:

$$\int_{\varphi^{-1}(Y_i)} \int_{X} \delta(X,X_i)k(X_i|Y_i,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X\,\mathrm{d}X_i = \int_{\varphi^{-1}(Y_i)} k(X_i|Y_i,\Theta)\log\big(f(X_i|\Theta')\big)\mathrm{d}X_i$$

As a result, the conditional expectation $Q(\Theta' \mid \Theta)$ given an observed sample $\mathcal{Y} = \{Y_1, Y_2,\ldots, Y_N\}$ and a set of associated random variables $\mathcal{X} = \{X_1, X_2,\ldots, X_N\}$ is specified as follows:

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N} \int_{\varphi^{-1}(Y_i)} k(X_i|Y_i,\Theta)\log\big(f(X_i|\Theta')\big)\mathrm{d}X_i$$

Note, all X_i (s) are iid and they are not existent in fact. Because all X_i are iid, let X be the random variable representing every X_i and the equation of $Q(\Theta' \mid \Theta)$ is re-written according to equation 2.10.

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N} \int_{\varphi^{-1}(Y_i)} k(X|Y_i,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X \tag{2.10}$$

The similar proof of equation 2.10 in case that X_i (s) are discrete is found in (Bilmes, 1998, p. 4). In case that $f(X \mid \Theta)$ and $k(X \mid Y_i, \Theta)$ belong to exponential family, equation 2.10 becomes equation 2.11 with an observed sample $\mathcal{Y} = \{Y_1, Y_2,\ldots, Y_N\}$.

$$Q(\Theta'|\Theta) = \left(\sum_{i=1}^{N} E\big(\log\big(b(X)\big)\big|Y_i,\Theta\big)\right) + \left((\Theta')^T \sum_{i=1}^{N} \tau_{\Theta,Y_i}\right) - N\log\big(a(\Theta')\big) \tag{2.11}$$

Where,

$$E\big(\log\big(b(X)\big)\big|Y_i,\Theta\big) = \int_{\varphi^{-1}(Y_i)} k(X|Y_i,\Theta)\log\big(b(X)\big)\mathrm{d}X$$

$$\tau_{\Theta,Y_i} = E(\tau(X)|Y_i,\Theta) = \int_{\varphi^{-1}(Y_i)} k(X|Y_i,\Theta)\tau(X)\mathrm{d}X$$

Please combine equation 2.9 and equation 2.10 to comprehend how to derive equation 2.11. Note, τ_{Θ,Y_i} is dependent on both Θ and Y_i.

DLR (Dempster, Laird, & Rubin, 1977, p. 1) called $\boldsymbol{X}$ as *complete data* because the mapping φ: $\boldsymbol{X} \to \boldsymbol{Y}$ is many-one function. There is another case that the complete space $\boldsymbol{Z}$ consists of hidden space $\boldsymbol{X}$ and observed space $\boldsymbol{Y}$ with note that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are separated. There is no explicit mapping φ from $\boldsymbol{X}$ and $\boldsymbol{Y}$ but there exists a PDF of $Z \in \boldsymbol{Z}$ as the joint PDF of $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$.

$$f(Z|\Theta) = f(X,Y|\Theta)$$

In this case, the equation 2.8 is modified with the joint PDF $f(X, Y \mid \Theta)$. The PDF of Y becomes:

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

The PDF $f(Y|\Theta)$ is equivalent to the PDF $g(Y|\Theta)$ mentioned in equation 1.34. Although there is no explicit mapping from $\boldsymbol{X}$ to $\boldsymbol{Y}$, the PDF of Y above implies an implicit mapping from $\boldsymbol{Z}$ to $\boldsymbol{Y}$. The conditional PDF of X given Z is specified according to Bayes' rule as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$f(Z|Y,\Theta) = f(X,Y|Y,\Theta) = f(X|Y)f(Y|Y) = f(X|Y,\Theta) = \frac{f(X,Y|\Theta)}{f(Y|\Theta)} = \frac{f(X,Y|\Theta)}{\int_X f(X,Y|\Theta)\mathrm{d}X}$$

The conditional PDF $f(X|Y, \Theta)$ is equivalent to the conditional PDF $k(X|Y, \Theta)$ mentioned in equation 1.35. Of course, given Y, we always have:

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

Equation 2.12 specifies the conditional expectation $Q(\Theta' \mid \Theta)$ in case that there is no explicit mapping from $\boldsymbol{X}$ to $\boldsymbol{Y}$ but there exists the joint PDF of X and Y.

$$Q(\Theta'|\Theta) = \int_{\boldsymbol{X}} f(Z|Y,\Theta)\log\big(f(Z|\Theta')\big)\mathrm{d}X = \int_{\boldsymbol{X}} f(X|Y,\Theta)\log\big(f(X,Y|\Theta')\big)\mathrm{d}X \qquad (2.12)$$

Where,

$$f(X|Y,\Theta) = \frac{f(X,Y|\Theta)}{f(Y|\Theta)} = \frac{f(X,Y|\Theta)}{\int_X f(X,Y|\Theta)\mathrm{d}X}$$

Note, $\boldsymbol{X}$ is separated from $\boldsymbol{Y}$ and the complete data $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$ is composed of $\boldsymbol{X}$ and $\boldsymbol{Y}$. For equation 2.12, the existence of the joint PDF $f(X, Y \mid \Theta)$ can be replaced by the existence of the conditional PDF $f(Y|X, \Theta)$ and the prior PDF $f(X|\Theta)$ due to:

$$f(X,Y|\Theta) = f(Y|X,\Theta)f(X|\Theta)$$

In applied statistics, equation 2.8 is often replaced by equation 2.12 because specifying the joint PDF $f(X, Y \mid \Theta)$ is more practical than specifying the mapping φ: $\boldsymbol{X} \rightarrow \boldsymbol{Y}$. However, equation 2.8 is more general equation 2.12 because the requirement of the joint PDF for equation 2.12 is stricter than the requirement of the explicit mapping for equation 2.8. In case that X and Y are discrete, equation 2.12 becomes:

$$Q(\Theta'|\Theta) = \sum_{X\in\boldsymbol{X}} P(X|Y,\Theta)\log\big(P(X,Y|\Theta')\big)$$

In case that X and Y are discrete, $P(X, Y \mid \Theta)$ is the joint probability of X and Y whereas $P(X \mid Y, \Theta)$ is the conditional probability of X given Y.

Equation 2.12 can be proved alternately without knowledge related to complete data (Sean, 2009). This proof is like the proof of equation 2.8. In fact, given hidden space $\boldsymbol{X}$, observed space $\boldsymbol{Y}$, and a joint PDF $f(X, Y \mid \Theta)$, the likelihood function $L(\Theta')$ is re-defined here as $\log(f(Y \mid \Theta'))$. The maximizer is:

$$\Theta^* = \underset{\Theta'}{\operatorname{argmax}}\, L(\Theta') = \underset{\Theta'}{\operatorname{argmax}} \log\big(f(Y|\Theta')\big)$$

Suppose the current parameter is Θ after some iteration. Next we must find out the new estimate Θ^* that maximizes the next log-likelihood function $L(\Theta')$. Suppose the total probability of observed data can be determined by marginalizing over hidden data:

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

The expansion of $f(Y \mid \Theta')$ is total probability rule. The next log-likelihood function $L(\Theta')$ is re-written:

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

Because hidden X is the complete set of mutually exclusive variables, the sum of conditional probabilities of X is equal to 1 given Y and Θ.

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

Where,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$f(X|Y,\Theta) = \frac{f(X,Y|\Theta)}{\int_X f(X,Y|\Theta)\mathrm{d}X}$$

Applying Jensen's inequality (Sean, 2009, pp. 3-4) with concavity of logarithm function

$$\log\left(\int_x u(x)v(x)\mathrm{d}x\right) \geq \int_x u(x)\log\big(v(x)\big)\mathrm{d}x$$

$$\text{where} \int_x u(x)\mathrm{d}x = 1$$

into $L(\Theta')$, we have (Sean, 2009, p. 6):

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

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

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

$$= Q(\Theta'|\Theta) - H(\Theta|\Theta)$$

Where,

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

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

Obviously, the lower-bound of $L(\Theta')$ is:

$$lb(\Theta'|\Theta) = Q(\Theta'|\Theta) - H(\Theta|\Theta)$$

As aforementioned, the lower-bound $lb(\Theta'|\Theta)$ (Sean, 2009, pp. 7-8) is maximized over many iterations of the iterative process so that $L(\Theta')$ is maximized finally. Because $H(\Theta|\Theta)$ is constant with regard to Θ', it is possible to eliminate $H(\Theta|\Theta)$ so that maximizing $Q(\Theta'|\Theta)$ is the same to maximizing the lower bound. In final, when GEM converges $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$, we have:

$$\Theta^* = \underset{\Theta'}{\operatorname{argmax}}\, lb(\Theta'|\Theta) = \underset{\Theta'}{\operatorname{argmax}}\, Q(\Theta'|\Theta)$$

We have the proof ■

Mixture model mentioned in subsection 5.1 is a good example for GEM without explicit mapping from $\boldsymbol{X}$ to $\boldsymbol{Y}$. Another well-known example is three-coin toss example (Collins & Barzilay, 2005) which applies GEM into estimating parameters of binomial distributions without explicit mapping.

Example 2.1. There are three coins named coin 1, coin 2 and coin 3. Each coin has two sides such as head (H) side and tail (T) side. Let hidden random variable X represent coin 1 where X is binary ($X = \{H, T\}$). Let θ_1 be probability of coin 1 receiving head side.

$$\theta_1 = P(X=H)$$

Of course, we have:

$$P(X=T) = 1 - \theta_1$$

Let observed random variable Y represent a sequence of tossing coin 2 or coin 3 three times. Such sequence depends on first tossing coin 1. For instance, if coin 1 shows head side ($X=H$), the sequence is result of tossing coin 2 three times. Otherwise, if coin 1 shows tail side ($X=T$), the sequence is result of tossing coin 3 three times. For example, suppose first tossing coin 1 results $X=H$ then, a possible result $Y = HHT$ means that we toss coin 2 three times resulting

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

head, head, and tail from coin 2. Obviously, *X* is hidden and *Y* is observed. In this example, we observe that

$$Y=HHT$$

Suppose *Y* conforms binomial distribution as follows:

$$P(Y|X)=\begin{cases}\theta_2^h(1-\theta_2)^t \text{ if } X=H\\ \theta_3^h(1-\theta_3)^t \text{ if } X=T\end{cases}$$

Where θ_2 and θ_3 are probabilities of coin 2 and coin 3 receiving head side, respectively. Note, *h* is the number of head side from trials of tossing coin 2 (if *X*=*H*) or coin 3 (if *X*=*T*). Similarly, *t* is the number of tail side from trials of tossing coin 2 (if *X*=*H*) or coin 3 (if *X*=*T*). The joint probability *P*(*X*, *Y*) is:

$$P(X,Y)=P(X)P(Y|X)=\begin{cases}\theta_1\theta_2^h(1-\theta_2)^t \text{ if } X=H\\ (1-\theta_1)\theta_3^h(1-\theta_3)^t \text{ if } X=T\end{cases}$$

In short, we need to estimate $\Theta=(\theta_1, \theta_2, \theta_3)^T$ from the observation *Y*=*HHT* by discrete version of $Q(\Theta' \mid \Theta)$. Given *Y*=*HHT*, we have *h*=2 and *t*=1. Thus, the probability *P*(*Y*|*X*) becomes:

$$P(Y|X)=P(Y=HHT|X)=\begin{cases}\theta_2^2(1-\theta_2) \text{ if } X=H\\ \theta_3^2(1-\theta_3) \text{ if } X=T\end{cases}$$

The joint probability *P*(*X*, *Y*) becomes:

$$P(X,Y)=\begin{cases}\theta_1\theta_2^2(1-\theta_2) \text{ if } X=H\\ (1-\theta_1)\theta_3^2(1-\theta_3) \text{ if } X=T\end{cases}$$

The probability of *Y* is calculated as follows:

$$P(Y)=P(Y|X=H)+P(Y|X=T)=\theta_2^2(1-\theta_2)+\theta_3^2(1-\theta_3)$$

The conditional probability of *X* given *Y* is determined as follows:

$$P(X|Y)=\frac{P(X,Y)}{P(Y)}=\begin{cases}\dfrac{\theta_1\theta_2^2(1-\theta_2)}{\theta_2^2(1-\theta_2)+\theta_3^2(1-\theta_3)} \text{ if } X=H\\ \dfrac{(1-\theta_1)\theta_3^2(1-\theta_3)}{\theta_2^2(1-\theta_2)+\theta_3^2(1-\theta_3)} \text{ if } X=T\end{cases}$$

The discrete version of $Q(\Theta' \mid \Theta)$ is determined as follows:

$$Q(\Theta'|\Theta)=\sum_{X\in\boldsymbol{X}}P(X|Y,\Theta)\log\big(P(X,Y|\Theta')\big)$$

$$=P(X=H|Y,\Theta)\log\big(P(X=H,Y|\Theta')\big)+P(X=T|Y,\Theta)\log\big(P(X=T,Y|\Theta')\big)$$

$$=\frac{\theta_1\theta_2^2(1-\theta_2)}{\theta_2^2(1-\theta_2)+\theta_3^2(1-\theta_3)}\log\big(\theta_1'(\theta_2')^2(1-\theta_2')\big)$$

$$+\frac{(1-\theta_1)\theta_3^2(1-\theta_3)}{\theta_2^2(1-\theta_2)+\theta_3^2(1-\theta_3)}\log\big((1-\theta_1')(\theta_3')^2(1-\theta_3')\big)$$

$$=\frac{\theta_1\theta_2^2(1-\theta_2)}{\theta_2^2(1-\theta_2)+\theta_3^2(1-\theta_3)}\big(\log(\theta_1')+2\log(\theta_2')+\log(1-\theta_2')\big)$$

$$+\frac{(1-\theta_1)\theta_3^2(1-\theta_3)}{\theta_2^2(1-\theta_2)+\theta_3^2(1-\theta_3)}\big(\log(1-\theta_1')+2\log(\theta_3')+\log(1-\theta_3')\big)$$

Note, $Q(\Theta'|\Theta)$ is function of $\Theta'=(\theta_1', \theta_2', \theta_3')^T$. The next parameter $\Theta^{(t+1)}=(\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_3^{(t+1)})^T$ is a maximizer of $Q(\Theta'|\Theta)$ with regard to Θ', which is solution of the equation created by setting the first-order derivative of $Q(\Theta'|\Theta)$ to be zero with note that the current parameter is $\Theta^{(t)}=\Theta$.

The first-order partial derivative of $Q(\Theta'|\Theta)$ with regard to θ_1' is:

$$\frac{\partial Q(\Theta'|\Theta)}{\partial\theta_1'}=\frac{\theta_1\theta_2^2(1-\theta_2)}{\theta_2^2(1-\theta_2)+\theta_3^2(1-\theta_3)}\frac{1}{\theta_1'}-\frac{(1-\theta_1)\theta_3^2(1-\theta_3)}{\theta_2^2(1-\theta_2)+\theta_3^2(1-\theta_3)}\frac{1}{1-\theta_1'}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$= \frac{\theta_1\theta_2^2(1-\theta_2) - \theta_1'\left(\theta_1\theta_2^2(1-\theta_2) + (1-\theta_1)\theta_3^2(1-\theta_3)\right)}{\theta_1'(1-\theta_1')\left(\theta_2^2(1-\theta_2) + \theta_3^2(1-\theta_3)\right)}$$

Setting this partial derivative $\frac{\partial Q(\Theta'|\Theta)}{\partial \theta_1'}$ to be zero, we obtain:

$$\theta_1' = \frac{\theta_1\theta_2^2(1-\theta_2)}{\theta_1\theta_2^2(1-\theta_2) + (1-\theta_1)\theta_3^2(1-\theta_3)}$$

Therefore, in M-step, given current parameter $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)})^T$, the next partial parameter $\theta_1^{(t+1)}$ is calculated as follows:

$$\theta_1^{(t+1)} = \frac{\theta_1^{(t)}\left(\theta_2^{(t)}\right)^2\left(1-\theta_2^{(t)}\right)}{\theta_1^{(t)}\left(\theta_2^{(t)}\right)^2\left(1-\theta_2^{(t)}\right) + \left(1-\theta_1^{(t)}\right)\left(\theta_3^{(t)}\right)^2\left(1-\theta_3^{(t)}\right)}$$

The first-order partial derivative of $Q(\Theta'|\Theta)$ with regard to θ_2' is:

$$\frac{\partial Q(\Theta'|\Theta)}{\partial \theta_2'} = \frac{\theta_1\theta_2^2(1-\theta_2)}{\theta_2^2(1-\theta_2) + \theta_3^2(1-\theta_3)} \frac{2-3\theta_2'}{\theta_2'(1-\theta_2')}$$

Setting this partial derivative $\frac{\partial Q(\Theta'|\Theta)}{\partial \theta_2'}$ to be zero, we obtain:

$$\theta_2' = \frac{2}{3}$$

Therefore, in M-step, given current parameter $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)})^T$, the next partial parameter $\theta_2^{(t+1)}$ is fixed:

$$\theta_2^{(t+1)} = \frac{2}{3}$$

The first-order partial derivative of $Q(\Theta'|\Theta)$ with regard to θ_3' is:

$$\frac{\partial Q(\Theta'|\Theta)}{\partial \theta_3'} = \frac{(1-\theta_1)\theta_3^2(1-\theta_3)}{\theta_2^2(1-\theta_2) + \theta_3^2(1-\theta_3)} \frac{2-3\theta_3'}{\theta_3'(1-\theta_3')}$$

Setting this partial derivative $\frac{\partial Q(\Theta'|\Theta)}{\partial \theta_3'}$ to be zero, we obtain:

$$\theta_3' = \frac{2}{3}$$

Therefore, in M-step, given current parameter $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)})^T$, the next partial parameter $\theta_3^{(t+1)}$ is fixed:

$$\theta_3^{(t+1)} = \frac{2}{3}$$

In short, in M-step of some t^{th} iteration, given current parameter $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)})^T$, only $\theta_1^{(t+1)}$ is updated whereas both $\theta_2^{(t+1)}$ and $\theta_3^{(t+1)}$ are fixed with observation Y=*HHT*.

$$\theta_1^{(t+1)} = \frac{\theta_1^{(t)}\left(\theta_2^{(t)}\right)^2\left(1-\theta_2^{(t)}\right)}{\theta_1^{(t)}\left(\theta_2^{(t)}\right)^2\left(1-\theta_2^{(t)}\right) + \left(1-\theta_1^{(t)}\right)\left(\theta_3^{(t)}\right)^2\left(1-\theta_3^{(t)}\right)}$$

$$\theta_2^{(t+1)} = \theta_3^{(t+1)} = \frac{2}{3}$$

For instance, let $\Theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)})^T$ be initialized arbitrarily as $\theta_1^{(0)} = \theta_2^{(0)} = \theta_3^{(0)} = 0.5$, at the first iteration, we obtain:

$$\theta_1^{(1)} = \frac{0.5 * (0.5)^2 * (1-0.5)}{0.5 * (0.5)^2 * (1-0.5) + (1-0.5) * (0.5)^2 * (1-0.5)} = 0.5$$

$$\theta_2^{(1)} = \theta_3^{(1)} = \frac{2}{3}$$

At the second iteration with current parameter $\Theta^{(1)} = (\theta_1^{(1)}=0.5, \theta_2^{(1)}=2/3, \theta_3^{(1)}=2/3)^T$, we obtain:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\theta_1^{(2)} = \frac{0.5 * \left(\frac{2}{3}\right)^2 * \left(1 - \frac{2}{3}\right)}{0.5 * \left(\frac{2}{3}\right)^2 * \left(1 - \frac{2}{3}\right) + (1 - 0.5) * \left(\frac{2}{3}\right)^2 * \left(1 - \frac{2}{3}\right)} = 0.5$$

$$\theta_2^{(2)} = \theta_3^{(2)} = \frac{2}{3}$$

As a result, GEM inside this example converges at the second iteration with final estimate $\Theta^{(1)} = \Theta^{(2)} = \Theta^* = (\theta_1^*=0.5, \theta_2^*=2/3, \theta_3^*=2/3)^T$ ■

In practice, suppose Y is observed as a sample $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$ of size N with note that all Y_i (s) are mutually independent and identically distributed (iid). The observed sample $\mathcal{Y}$ is associated with a a hidden set (latent set) $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ of size N. All X_i (s) are iid and they are not existent in fact. Let $X \in \boldsymbol{X}$ be the random variable representing every X_i. Of course, the domain of X is $\boldsymbol{X}$. Equation 2.13 specifies the conditional expectation $Q(\Theta' \mid \Theta)$ given such $\mathcal{Y}$.

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N} \int_{\boldsymbol{X}} f(X|Y_i, \Theta) \log\left(f(X, Y_i|\Theta')\right) \mathrm{d}X \tag{2.13}$$

Equation 2.13 is a variant of equation 2.10 in case that there is no explicit mapping between X_i and Y_i but there exists the same joint PDF between X_i and Y_i. Please see the proof of equation 2.10 to comprehend how to derive equation 2.13. If both X and Y are discrete, equation 2.13 becomes:

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N} \sum_{X \in \boldsymbol{X}} P(X|Y_i, \Theta) \log\left(P(X, Y_i|\Theta')\right) \tag{2.14}$$

If X is discrete and Y is continuous such that $f(X, Y \mid \Theta) = P(X|\Theta) f(Y \mid X, \Theta)$ then, according to the total probability rule, we have:

$$f(Y|\Theta) = \sum_{X \in \boldsymbol{X}} P(X|\Theta) f(Y|X, \Theta)$$

Note, when only X is discrete, its PDF $f(X|\Theta)$ becomes the probability $P(X|\Theta)$. Therefore, equation 2.15 is a variant of equation 2.13, as follows:

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N} \sum_{X \in \boldsymbol{X}} P(X|Y_i, \Theta) \log\left(P(X|\Theta') f(Y_i|X, \Theta')\right) \tag{2.15}$$

Where $P(X \mid Y_i, \Theta)$ is determined by Bayes' rule, as follows:

$$P(X|Y_i, \Theta) = \frac{P(X|\Theta) f(Y_i|X, \Theta)}{\sum_X P(X|\Theta) f(Y_i|X, \Theta)}$$

Equation 2.15 is the base for estimating the probabilistic mixture model by EM algorithm, which will be described later in detail. Convergence of GEM will be mentioned in next section.

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:

$$L(\Theta) = \log\left(g(Y|\Theta)\right) = \log\left(\int_{\varphi^{-1}(Y)} f(X|\Theta) \mathrm{d}X\right)$$

$$Q(\Theta'|\Theta) = E\left(\log\left(f(X|\Theta')\right)\middle|Y, \Theta\right) = \int_{\varphi^{-1}(Y)} k(X|Y, \Theta) \log\left(f(X|\Theta')\right) \mathrm{d}X$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

$$H(\Theta'|\Theta) = E\big(\log\big(k(X|Y,\Theta')\big)\big|Y,\Theta\big) = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta')\big)\mathrm{d}X \tag{3.1}$$

From equation 2.8 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.

$$Q(\Theta'|\Theta) = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(g(Y|\Theta')k(X|Y,\Theta')\big)\mathrm{d}X$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(g(Y|\Theta')\big)\mathrm{d}X + \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta')\big)\mathrm{d}X$$

$$= \log\big(g(Y|\Theta')\big)\int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\mathrm{d}X + H(\Theta'|\Theta) = \log\big(g(Y|\Theta')\big) + H(\Theta'|\Theta)$$

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

Lemma 3.1 (Dempster, Laird, & Rubin, 1977, p. 6). For any pair (Θ', Θ) in Ω x Ω,

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

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

Following is a proof of lemma 3.1 as well as equation 3.3. The log-likelihood function $L(\Theta')$ is re-written as follows:

$$L(\Theta') = \log\left(\int_{\varphi^{-1}(Y)} f(X|\Theta')\mathrm{d}X\right) = \log\left(\int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\frac{f(X|\Theta')}{k(X|Y,\Theta)}\mathrm{d}X\right)$$

Due to

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

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

$$\log\left(\int_x u(x)v(x)\mathrm{d}x\right) \ge \int_x u(x)\log\big(v(x)\big)\mathrm{d}x$$

$$\text{where}\int_x u(x)\mathrm{d}x = 1$$

into $L(\Theta')$, we have (Sean, 2009, p. 6):

$$L(\Theta') \ge \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\left(\frac{f(X|\Theta')}{k(X|Y,\Theta)}\right)\mathrm{d}X$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\left(\log\big(f(X|\Theta')\big) - \log\big(k(X|Y,\Theta)\big)\right)\mathrm{d}X$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X - \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta)\big)\mathrm{d}X$$

$$= Q(\Theta'|\Theta) - H(\Theta|\Theta)$$

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

(Due to $Q(\Theta'|\Theta) = L(\Theta') + H(\Theta'|\Theta)$)

It implies:

$$H(\Theta'|\Theta) \le H(\Theta|\Theta)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

According to Jensen's inequality (Sean, 2009, pp. 3-4), the equality $H(\Theta'|\Theta) = H(\Theta|\Theta)$ 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 and $k(X \mid Y, \Theta')$ is a PDF ■

We also have the lower-bound of $L(\Theta')$, denoted $lb(\Theta'|\Theta)$ as follows:

$$lb(\Theta'|\Theta) = Q(\Theta'|\Theta) - H(\Theta|\Theta)$$

Obviously, we have:

$$L(\Theta') \geq lb(\Theta'|\Theta)$$

As aforementioned, the lower-bound $lb(\Theta'|\Theta)$ is maximized over many iterations of the iterative process so that $L(\Theta')$ is maximized finally. Such lower-bound is determined indirectly by $Q(\Theta'|\Theta)$ so that maximizing $Q(\Theta'|\Theta)$ with regard to Θ' is the same to maximizing $lb(\Theta'|\Theta)$ because $H(\Theta|\Theta)$ is constant with regard to Θ'.

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

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

Definition 3.1 (Dempster, Laird, & Rubin, 1977, p. 7). An iterative algorithm with mapping $M(\Theta)$ is a GEM algorithm if

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

Of course, specification of GEM shown in table 2.3 satisfies the definition 3.1 because $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with regard to variable Θ in M-step.

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

Theorem 3.1 (Dempster, Laird, & Rubin, 1977, p. 7). For every GEM algorithm

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

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 3.1 (Dempster, Laird, & Rubin, 1977, p. 7):

$$\begin{aligned} L\left(M(\Theta)\right) - L(\Theta) &= \left(Q(M(\Theta)|\Theta) - H(M(\Theta)|\Theta)\right) - \left(Q(\Theta|\Theta) - H(\Theta|\Theta)\right) \\ &= \left(Q(M(\Theta)|\Theta) - Q(\Theta|\Theta)\right) + \left(H(\Theta|\Theta) - H(M(\Theta)|\Theta)\right) \geq 0 \blacksquare \end{aligned}$$

Because the equality of lemma 3.1 occurs if and only if $k(X \mid Y, \Theta') = k(X \mid Y, \Theta)$ almost everywhere and the equality of the definition 3.1 is $Q(M(\Theta) \mid \Theta) = Q(\Theta \mid \Theta)$, we deduce that the equality of theorem 3.1 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 3.1 and corollary 3.2 from definition 3.1 and theorem 3.1.

Corollary 3.1 (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:

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

(2) $Q(M(\Theta^*) \mid \Theta^*) = Q(\Theta^* \mid \Theta^*)$

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

Proof. From theorem 3.1 and the assumption of corollary 3.1, we have:

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

This implies:

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

As a result,

$$L\left(M(\Theta^*)\right) = L(\Theta^*)$$

From theorem 3.1, we also have:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

Corollary 3.2 (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 3.1 and the assumption of corollary 3.2, we have:

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

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

Theorem 3.2 (Dempster, Laird, & Rubin, 1977, p. 7). Suppose $\left\{\Theta^{(t)}\right\}_{t=1}^{+\infty}$ is the sequence of estimates resulted from GEM algorithm such that:

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

Proof. The sequence $\left\{L\left(\Theta^{(t)}\right)\right\}_{t=1}^{+\infty}$ is non-decreasing according to theorem 3.1 and is bounded above according to the assumption 1 of theorem 3.2 and hence, the sequence $\left\{L\left(\Theta^{(t)}\right)\right\}_{t=1}^{+\infty}$ 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 $v \geq 1$:

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

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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:

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

According to Cauchy criterion, the sequence $\left\{\Theta^{(t)}\right\}_{t=1}^{+\infty}$ converges to some Θ^* in the closure of Ω.

Theorem 3.1 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 3.1 and 3.2 indicate that the optimal estimate is a fixed point of GEM algorithm. Theorem 3.2 points out convergence condition of GEM algorithm but does not assert the converged point Θ^* is maximizer of $L(\Theta)$. So, we need mathematical tools of derivative and differential to prove convergence of GEM to a maximizer Θ^*. 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^{th}-order partial derivative (differential) with regard to first variable and then, taking j^{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 $D^{11}Q(\Theta'|\Theta) = \frac{\partial^2 Q(\Theta'|\Theta)}{\partial\Theta'\partial\Theta}$
- Secondly, we substitute $\Theta^{(t)}$ and $\Theta^{(t+1)}$ for such $D^{11}Q(\Theta' \mid \Theta)$ to obtain $D^{11}Q(\Theta^{(t)} \mid \Theta^{(t+1)})$.

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

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

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

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

Lemma 3.2 (Dempster, Laird, & Rubin, 1977, p. 8). For all Θ in Ω,

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

$$\begin{aligned} V_N\left(\frac{\mathrm{dlog}\big(k(X|Y,\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right) &= E\left(\left(\frac{\mathrm{dlog}\big(k(X|Y,\Theta)\big)}{\mathrm{d}\Theta}\right)^2\middle|Y,\Theta\right) \\ &= -E\left(\frac{d^2\log\big(k(X|Y,\Theta)\big)}{\mathrm{d}(\Theta)^2}\middle|Y,\Theta\right) \end{aligned} \tag{3.9}$$

$$D^{10}Q(\Theta|\Theta) = DL(\Theta) = E\left(\frac{\mathrm{dlog}\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right) \tag{3.10}$$

$$D^{20}Q(\Theta|\Theta) = D^2L(\Theta) + D^{20}H(\Theta|\Theta) = E\left(\frac{d^2\log\big(f(X|\Theta)\big)}{\mathrm{d}(\Theta)^2}\middle|Y,\Theta\right) \tag{3.11}$$

$$\begin{aligned} V_N\left(\frac{\mathrm{dlog}\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right) &= E\left(\left(\frac{\mathrm{dlog}\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\right)^2\middle|Y,\Theta\right) \\ &= D^2L(\Theta) + \big(DL(\Theta)\big)^2 - D^{20}Q(\Theta|\Theta)\blacksquare \end{aligned} \tag{3.12}$$

Note, $V_N(.)$ denotes non-central variance (non-central covariance matrix). Followings are proofs of equation 3.7, equation 3.8, equation 3.9, equation 3.10, equation 3.11, and equation 3.12. In fact, we have:

$$D^{10}H(\Theta'|\Theta) = \frac{\partial}{\partial\Theta'}E\big(\log\big(k(X|Y,\Theta')\big)\big|Y,\Theta\big) = \frac{\partial}{\partial\Theta'}\left(\int\limits_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\big(k(X|Y,\Theta')\big)\mathrm{d}X\right)$$

$$= \int\limits_{\varphi^{-1}(Y)} k(X|Y,\Theta)\frac{\mathrm{dlog}\big(k(X|Y,\Theta')\big)}{\mathrm{d}\Theta'}\mathrm{d}X = E\left(\frac{\mathrm{dlog}\big(k(X|Y,\Theta')\big)}{\mathrm{d}\Theta'}\middle|Y,\Theta\right) =$$

$$= \int\limits_{\varphi^{-1}(Y)} \frac{k(X|Y,\Theta)}{k(X|Y,\Theta')}\frac{\mathrm{d}\big(k(X|Y,\Theta')\big)}{\mathrm{d}\Theta'}\mathrm{d}X$$

It implies:

$$D^{10}H(\Theta|\Theta) = \int\limits_{\varphi^{-1}(Y)} \frac{k(X|Y,\Theta)}{k(X|Y,\Theta)}\frac{\mathrm{d}\big(k(X|Y,\Theta)\big)}{\mathrm{d}\Theta}\mathrm{d}X = \frac{\mathrm{d}}{\mathrm{d}\Theta}\left(\int\limits_{\varphi^{-1}(Y)} k(X|Y,\Theta)\mathrm{d}X\right) = \frac{\mathrm{d}}{\mathrm{d}\Theta}(1)$$
$$= \mathbf{0}^T$$

Thus, equation 3.7 is proved.

We also have:

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

It implies:

$$D^{11}H(\Theta|\Theta) = \int\limits_{\varphi^{-1}(Y)} \frac{1}{k(X|Y,\Theta)}\frac{\mathrm{d}k(X|Y,\Theta)}{d\Theta}\frac{\mathrm{d}k(X|Y,\Theta)}{\mathrm{d}\Theta}\mathrm{d}X$$

$$= \int\limits_{\varphi^{-1}(Y)} k(X|Y,\Theta)\left(\frac{1}{k(X|Y,\Theta)}\frac{\mathrm{d}k(X|Y,\Theta)}{d\Theta}\right)^2\mathrm{d}X = V_N\left(\frac{\mathrm{dlog}\big(k(X|Y,\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

We also have:

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

$$= -\int_{\varphi^{-1}(Y)} \frac{k(X|Y,\Theta)}{\left(k(X|Y,\Theta')\right)^2}\left(\frac{\mathrm{d}k(X|Y,\Theta')}{\mathrm{d}\Theta'}\right)^2 \mathrm{d}X = -E\left(\left(\frac{\mathrm{d}\log\left(k(X|Y,\Theta')\right)}{\mathrm{d}\Theta'}\right)^2\middle|Y,\Theta\right)$$

It implies:

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

$$= -V_N\left(\frac{\mathrm{d}\log\left(k(X|Y,\Theta)\right)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)$$

Hence, equation 3.8 and equation 3.9 are proved.

From equation 3.2, we have:

$$D^{20}Q(\Theta'|\Theta) = D^2L(\Theta') + D^{20}H(\Theta'|\Theta)$$

We also have:

$$D^{10}Q(\Theta'|\Theta) = \frac{\partial}{\partial\Theta'}\left(\int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\left(f(X|\Theta')\right)\mathrm{d}X\right)$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\frac{\mathrm{d}\log\left(f(X|\Theta')\right)}{\mathrm{d}\Theta'}\mathrm{d}X$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\frac{\mathrm{d}\log\left(f(X|\Theta')\right)}{\mathrm{d}\Theta'}\mathrm{d}X = E\left(\frac{\mathrm{d}\log\left(f(X|\Theta')\right)}{\mathrm{d}\Theta'}\middle|Y,\Theta\right)$$

$$= \int_{\varphi^{-1}(Y)} \frac{k(X|Y,\Theta)}{f(X|\Theta')}\frac{\mathrm{d}f(X|\Theta')}{\mathrm{d}\Theta'}\mathrm{d}X$$

It implies:

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

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

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

Thus, equation 3.10 is proved.

We have:

$$D^{20}Q(\Theta'|\Theta) = \frac{\partial D^{10}Q(\Theta'|\Theta)}{\partial\Theta'} = \frac{\partial}{\partial\Theta'}\left(\int_{\varphi^{-1}(Y)} \frac{k(X|Y,\Theta)}{f(X|\Theta')}\frac{\mathrm{d}f(X|\Theta')}{\mathrm{d}\Theta'}\mathrm{d}X\right)$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\frac{d}{\mathrm{d}\Theta'}\left(\frac{\mathrm{d}f(X|\Theta')/\mathrm{d}\Theta'}{f(X|\Theta')}\right)\mathrm{d}X = E\left(\frac{\mathrm{d}^2\log\left(f(X|\Theta')\right)}{\mathrm{d}(\Theta')^2}\middle|Y,\Theta\right)$$

(Hence, equation 3.11 is proved)

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\left(\left(\mathrm{d}^2f(X|\Theta')/\mathrm{d}(\Theta')^2\right)f(X|\Theta') - \left(\mathrm{d}f(X|\Theta')/\mathrm{d}\Theta'\right)^2\right)/\left(f(X|\Theta')\right)^2\mathrm{d}X$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\frac{(\mathrm{d}^2 f(X|\Theta')/\mathrm{d}(\Theta')^2)}{f(X|\Theta')}\mathrm{d}X - \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\left(\frac{\mathrm{d}f(X|\Theta')/\mathrm{d}\Theta'}{f(X|\Theta')}\right)^2 \mathrm{d}X$$

$$= \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\frac{(\mathrm{d}^2 f(X|\Theta')/\mathrm{d}(\Theta')^2)}{f(X|\Theta')}\mathrm{d}X - V_N\left(\frac{\mathrm{dlog}\big(f(X|\Theta')\big)}{\mathrm{d}\Theta'}\middle|Y,\Theta\right)$$

It implies:

$$D^{20}Q(\Theta|\Theta) = \int_{\varphi^{-1}(Y)} k(X|Y,\Theta)\frac{(\mathrm{d}^2 f(X|\Theta)/\mathrm{d}(\Theta)^2)}{f(X|\Theta)}\mathrm{d}X - V_N\left(\frac{\mathrm{dlog}\big(f(X|\Theta)\big)}{\mathrm{d}\Theta}\middle|Y,\Theta\right)$$

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

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

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

Due to:

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

We have:

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

Therefore, equation 3.12 is proved ■

Lemma 3.3 (Dempster, Laird, & Rubin, 1977, p. 9). If *f*(*X* | Θ) and *k*(*X* | *Y*, Θ) belong to exponential family, for all Θ in Ω, we have:

$$D^{10}H(\Theta'|\Theta) = E(\tau(X)|Y,\Theta) - E(\tau(X)|Y,\Theta') \quad (3.13)$$

$$D^{20}H(\Theta'|\Theta) = -V(\tau(X)|Y,\Theta') \quad (3.14)$$

$$D^{10}Q(\Theta'|\Theta) = E(\tau(X)|\Theta) - E(\tau(X)|\Theta') \quad (3.15)$$

$$D^{20}Q(\Theta'|\Theta) = -V(\tau(X)|\Theta')\blacksquare \quad (3.16)$$

Proof. If *f*(*X* | Θ') and *k*(*X* | *Y*, Θ') belong to exponential family, from table 1.2 we have:

$$\frac{\mathrm{dlog}\big(f(Y|\Theta')\big)}{\mathrm{d}\Theta'} = \frac{\mathrm{d}}{\mathrm{d}\Theta'}\big(b(X)\exp\big((\Theta')^T\tau(X)\big)/a(\Theta')\big) = \tau(X) - \mathrm{log}'\big(a(\Theta')\big)$$
$$= \tau(X) - E(\tau(X)|\Theta')$$

And,

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

And,

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

And,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

Hence,

$$D^{10}H(\Theta'|\Theta)=\frac{\partial}{\partial\Theta'}\left(\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\log\left(k(X|Y,\Theta')\right)\mathrm{d}X\right)$$
$$=\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\frac{\mathrm{d}\log\left(k(X|Y,\Theta')\right)}{\mathrm{d}\Theta'}\mathrm{d}X$$
$$=\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\tau(X)\mathrm{d}X-\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)E(\tau(X)|Y,\Theta')\mathrm{d}X$$
$$=E(\tau(X)|Y,\Theta)-E(\tau(X)|Y,\Theta')\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\mathrm{d}X=E(\tau(X)|Y,\Theta)-E(\tau(X)|Y,\Theta')$$

We have:

$$D^{20}H(\Theta'|\Theta)=\frac{\partial^2}{\partial(\Theta')^2}\left(\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\log\left(k(X|Y,\Theta')\right)\mathrm{d}X\right)$$
$$=\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\frac{\mathrm{d}^2\log\left(k(X|Y,\Theta')\right)}{\mathrm{d}(\Theta')^2}\mathrm{d}X=-\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\log''(a(\Theta')|Y)\mathrm{d}X$$
$$=-\log''(a(\Theta')|Y)\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\mathrm{d}X=-\log''(a(\Theta')|Y)=-V(\tau(X)|Y,\Theta')$$

We have:

$$D^{10}Q(\Theta'|\Theta)=\frac{\partial}{\partial\Theta'}\left(\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\log\left(f(X|\Theta')\right)\mathrm{d}X\right)$$
$$=\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\frac{\mathrm{d}\log\left(f(X|\Theta')\right)}{\mathrm{d}\Theta'}\mathrm{d}X$$
$$=\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\tau(X)\mathrm{d}X-\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)E(\tau(X)|\Theta)\mathrm{d}X$$
$$=E(\tau(X)|\Theta)-E(\tau(X)|\Theta')\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\mathrm{d}X=E(\tau(X)|\Theta)-E(\tau(X)|\Theta')$$

We have:

$$D^{20}Q(\Theta'|\Theta)=\frac{\partial^2}{\partial(\Theta')^2}\left(\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\log\left(f(X|\Theta')\right)\mathrm{d}X\right)$$
$$=\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\frac{\mathrm{d}^2\log\left(f(X|\Theta')\right)}{\mathrm{d}(\Theta')^2}\mathrm{d}X=-\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\log''\left(a(\Theta')\right)\mathrm{d}X$$
$$=-\log''\left(a(\Theta')\right)\int\limits_{\varphi^{-1}(Y)}k(X|Y,\Theta)\mathrm{d}X=-\log''\left(a(\Theta')\right)=-V(\tau(X)|\Theta')\blacksquare$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

Theorem 3.3 (Dempster, Laird, & Rubin, 1977, p. 8). Suppose the sequence $\{\Theta^{(t)}\}_{t=1}^{+\infty}$ is an instance of GEM algorithm such that

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

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

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

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 Θ^* in the closure of Ω ■

Note, if Θ 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 3.3.

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

$$\begin{aligned} Q\left(\Theta\middle|\Theta^{(t)}\right) &= Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) + D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta - \Theta^{(t+1)}\right) \\ &\quad + \left(\Theta - \Theta^{(t+1)}\right)^T D^{20}Q\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta - \Theta^{(t+1)}\right) \\ &= Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) + \left(\Theta - \Theta^{(t+1)}\right)^T D^{20}Q\left(\Theta_0^{(t+1)}\middle|\Theta^{(t)}\right)\left(\Theta - \Theta^{(t+1)}\right) \\ &\quad \left(\text{due to } D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) = \mathbf{0}^T\right) \end{aligned}$$

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

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

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

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

Whereas,

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

So, for all t, there exists some $\xi > 0$ such that

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

In other words, the assumption 2 of theorem 3.2 is satisfied and hence, the sequence $\{\Theta^{(t)}\}_{t=1}^{+\infty}$ converges to some Θ^* in the closure of Ω if the sequence $\{L(\Theta^{(t)})\}_{t=1}^{+\infty}$ is bounded above ■

Theorem 3.4 (Dempster, Laird, & Rubin, 1977, p. 9). Suppose the sequence $\{\Theta^{(t)}\}_{t=1}^{+\infty}$ is an instance of GEM algorithm such that

(1) The sequence $\{\Theta^{(t)}\}_{t=1}^{+\infty}$ converges to Θ^* in the closure of Ω.
(2) $D^{10}Q(\Theta^{(t+1)} \mid \Theta^{(t)}) = \mathbf{0}^T$ for all t.
(3) $D^{20}Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is negative definite for all t.

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

$$DM(\Theta^*) = D^{20}H(\Theta^*|\Theta^*)\left(D^{20}Q(\Theta^*|\Theta^*)\right)^{-1} \blacksquare \tag{3.17}$$

The notation "-1" denotes inverse of matrix. Note, $DM(\Theta^*)$ is differential of $M(\Theta)$ at $\Theta = \Theta^*$, which implies convergence rate of GEM algorithm. Obviously, Θ^* is local maximizer due to $DL(\Theta^*) = \mathbf{0}^T$ and $D^{20}Q(\Theta^* \mid \Theta^*)$. Followings are proofs of theorem 3.4.

From equation 3.2, we have:

$$\begin{aligned} DL\left(\Theta^{(t+1)}\right) &= D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - D^{10}H\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) = -D^{10}H\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) \\ &\quad \left(\text{Due to } D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) = \mathbf{0}^T\right) \end{aligned}$$

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 $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

By first-order Taylor series expansion for $D^{10}Q(\Theta_2 \mid \Theta_1)$ as a function of Θ_1 at $\Theta_1 = \Theta^*$ and as a function of Θ_2 at $\Theta_2 = \Theta^*$, respectively, we have:

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

Where $R_1(\Theta_1)$ and $R_2(\Theta_2)$ are remainders. By summing such two series, we have:

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

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

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

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

$$\mathbf{0}^T = D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^*\right) + D^{10}Q\left(\Theta^*\middle|\Theta^{(t)}\right) + \left(\Theta^{(t)} - \Theta^*\right)^T D^{11}Q\left(\Theta^{(t+1)}\middle|\Theta^*\right) + \left(\Theta^{(t+1)} - \Theta^*\right)^T D^{20}Q\left(\Theta^*\middle|\Theta^{(t)}\right) + R_1\left(\Theta^{(t)}\right) + R_2\left(\Theta^{(t+1)}\right)$$

It implies:

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

Multiplying two sides of the equation above by $D^{20}Q(\Theta^* \mid \Theta^{(t)})^{-1}$ and letting $M(\Theta^{(t)}) = \Theta^{(t+1)}$, $M(\Theta^*) = \Theta^*$, we obtain:

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

Let t approach $+\infty$ such that $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$, we obtain $DM(\Theta^*)$ as differential of $M(\Theta)$ at Θ^* as follows:

$$DM(\Theta^*) = -D^{11}Q(\Theta^*|\Theta^*)\left(D^{20}Q(\Theta^*|\Theta^*)\right)^{-1} \quad (3.18)$$

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

$$D^{11}Q\left(\Theta^{(t+1)}\middle|\Theta^*\right) = D^{11}Q(\Theta^*|\Theta^*)$$
$$D^{20}Q\left(\Theta^*\middle|\Theta^{(t)}\right) = D^{20}Q(\Theta^*|\Theta^*)$$
$$D^{10}Q\left(\Theta^{(t+1)}\middle|\Theta^*\right) = D^{10}Q(\Theta^*|\Theta^*) = \mathbf{0}^T$$
$$D^{10}Q\left(\Theta^*\middle|\Theta^{(t)}\right) = D^{10}Q(\Theta^*|\Theta^*) = \mathbf{0}^T$$
$$\lim_{t\to+\infty} R_1\left(\Theta^{(t)}\right) = \lim_{\Theta^{(t)}\to\Theta^*} R_1\left(\Theta^{(t)}\right) = 0$$
$$\lim_{t\to+\infty} R_2\left(\Theta^{(t+1)}\right) = \lim_{\Theta^{(t+1)}\to\Theta^*} R_2\left(\Theta^{(t+1)}\right) = 0$$

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:

$$D^{11}Q(\Theta^*|\Theta^*) = DL(\Theta^*) + D^{11}H(\Theta^*|\Theta^*) = 0 + D^{11}H(\Theta^*|\Theta^*)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

(Due to theorem 3.4)

$= -D^{20}H(\Theta^*|\Theta^*)$

(Due to equation 3.8)

Therefore, equation 3.18 becomes equation 3.17.

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

Finally, theorem 3.4 is proved. By combination of theorems 3.2 and 3.4, I propose corollary 3.3 as a convergence criterion to local maximizer of GEM.

Corollary 3.3. 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) The sequence $\left\{L\left(\Theta^{(t)}\right)\right\}_{t=1}^{+\infty}$ is bounded above.

(3) $D^{10}Q(\Theta^* \mid \Theta^*) = \mathbf{0}^T$ and $D^{20}Q(\Theta^* \mid \Theta^*)$ negative definite with suppose that Θ^* is the converged point.

Then,

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

(2) Equation 3.17 is obtained ■

The assumption 1 of corollary 3.3 implies that the given algorithm is a GEM according to definition 3.1. From such assumption, we also have:

$$\begin{cases} Q\left(\Theta^{(t+1)}\middle|\Theta^{(t)}\right) - Q\left(\Theta^{(t)}\middle|\Theta^{(t)}\right) > 0 \\ \left(\Theta^{(t+1)} - \Theta^{(t)}\right)^T\left(\Theta^{(t+1)} - \Theta^{(t)}\right) \geq 0 \end{cases}$$

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

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

In other words, the assumption 2 of theorem 3.2 is satisfied and hence, the sequence $\left\{\Theta^{(t)}\right\}_{t=1}^{+\infty}$ converges to some Θ^* in the closure of Ω when the sequence $\left\{L\left(\Theta^{(t)}\right)\right\}_{t=1}^{+\infty}$ is bounded above according to the assumption 2 of corollary 3.3. From equation 3.2, we have:

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

When t approaches $+\infty$ such that $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^*$ then,

$$DL(\Theta^*) = D^{10}Q(\Theta^* \mid \Theta^*) - D^{10}H(\Theta^* \mid \Theta^*)$$

$D^{10}H(\Theta^* \mid \Theta^*)$ is zero according to equation 3.7. Hence, along with the assumption 3 of corollary 3.3, we have:

$$DL(\Theta^*) = D^{10}Q(\Theta^* \mid \Theta^*) = \mathbf{0}^T$$

Due to $DL(\Theta^*) = 0$, we only assert here that the given algorithm converges to Θ^* as a stationary point of $L(\Theta)$. Later on, we will prove that Θ^* is a local maximizer of $L(\Theta)$ when $Q(M(\Theta^{(t)}) \mid \Theta^{(t)}) > Q(\Theta^{(t)} \mid \Theta^{(t)})$, $DL(\Theta^*) = 0$, and $D^{20}Q(\Theta^* \mid \Theta^*)$ negative definite. Due to $D^{10}Q(\Theta^* \mid \Theta^*) = \mathbf{0}^T$, we obtain equation 3.17. Please see the proof of equation 3.17 ■

By default, suppose all GEM algorithms satisfy the assumptions 2 and 3 of corollary 3.3. Thus, we only check the assumption 1 to verify whether a given algorithm is a GEM which converges to local maximizer Θ^*. Note, if the assumption 1 of corollary 3.3 is replaced by "$Q(M(\Theta^{(t)}) \mid \Theta^{(t)}) \geq Q(\Theta^{(t)} \mid \Theta^{(t)})$ for all t" then, Θ^* is only asserted to be a stationary point of $L(\Theta)$ such that $DL(\Theta^*) = \mathbf{0}^T$. Wu (Wu, 1983) gave a deep research on convergence of GEM in her/his article "On the Convergence Properties of the EM Algorithm". Please read this article for more details about convergence of GEM.

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:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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 an (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 equation 3.19 and equation 3.20. Of course, h_1^*, h_2^*,…, h_r^* are eigenvalues of $D^{20}H(\Theta^* \mid \Theta^*)$ whereas q_1^*, q_2^*,…, q_r^* are eigenvalues of $D^{20}Q(\Theta^* \mid \Theta^*)$.

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

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

From equation 3.17, $DM(\Theta^*)$ is decomposed as seen in equation 3.21.

$$DM(\Theta^*) = (UH_e^*U^{-1})(UQ_e^*U^{-1})^{-1} = UH_e^*U^{-1}U(Q_e^*)^{-1}U^{-1} = U(H_e^*(Q_e^*)^{-1})U^{-1} \tag{3.21}$$

Let M_e^* be eigenvalue matrix of $DM(\Theta^*)$, specified by equation 3.17. As a convention M_e^* is called convergence matrix.

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

Of course, all $m_i^* = h_i^* / q_i^*$ are eigenvalues of $DM(\Theta^*)$ with assumption $q_i^* < 0$ 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 Θ degrades into scalar as $\Theta = \theta$ with note that scalar is 1-element vector, equation 3.17 is re-written as equation 3.23:

$$DM(\theta^*) = M_e^* = m^* = \lim_{t \to +\infty} \frac{M\left(\theta^{(t)}\right) - M(\theta^*)}{\theta^{(t)} - \theta^*} = \lim_{t \to +\infty} \frac{\theta^{(t+1)} - \theta^*}{\theta^{(t)} - \theta^*} = \\ = D^{20}H(\theta^*|\theta^*)\left(D^{20}Q(\theta^*|\theta^*)\right)^{-1} \tag{3.23}$$

From equation 3.23, the next estimate $\theta^{(t+1)}$ approaches θ^* when $t \to +\infty$ and so we have:

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

So equation 3.24 is a variant of equation 3.23 (McLachlan & Krishnan, 1997, p. 120).

$$DM(\theta^*) = M_e = m^* = \lim_{t \to +\infty} \frac{\theta^{(t+2)} - \theta^{(t+1)}}{\theta^{(t+1)} - \theta^{(t)}} \tag{3.24}$$

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:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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 contradicts the converse assumption "there always exists $m_i^* > 1$ or $m_i^* < 0$ for some i". Therefore, we conclude that $0 \leq m_i^* \leq 1$ for all i. In general, if Θ^* 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 3.25.

$$\begin{aligned} & q_i^* < 0, \forall i \\ & h_i^* \leq 0, \forall i \\ & 0 \leq m_i^* \leq 1, \forall i \end{aligned} \tag{3.25}$$

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^{th} iteration and $\Theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_r^*)$, each m_i^* measures how much the next $\theta_i^{(t+1)}$ is near to θ_i^*. In other words, the smaller the m_i^* (s) are, the faster the GEM is and so the better the GEM is. This is why DLR (Dempster, Laird, & Rubin, 1977, p. 10) defined that the convergence rate m^* of GEM is the maximum one among all m_i^*, as seen in equation 3.26. The convergence rate m^* implies lowest speed.

$$m^* = \max_{m_i^*} \{m_1^*, m_2^*, \ldots, m_r^*\} \text{ where } m_1^* = \frac{h_1^*}{q_1^*} \tag{3.26}$$

From equation 3.2 and equation 3.17, we have (Dempster, Laird, & Rubin, 1977, p. 10):

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

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:

$$L_e = (I - M_e)Q_e \tag{3.27}$$

Where L_e^* is eigenvalue matrix of $D^2L(\Theta^*)$. From equation 3.27, 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 3.28 specifies a so-called speed matrix S_e^*:

$$S_e^* = \begin{pmatrix} s_1^* = 1 - m_1^* & 0 & \cdots & 0 \\ 0 & s_2^* = 1 - m_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_r^* = 1 - m_r^* \end{pmatrix} \tag{3.28}$$

This implies

$$L_e^* = S_e^* Q_e^*$$

From equation 3.25 and equation 3.28, we have $0 \leq s_i^* \leq 1$. Equation 3.29 specifies L_e^* which is eigenvalue matrix of $D^2L(\Theta^*)$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$L_e^* = \begin{pmatrix} l_1^* = s_1^* q_1^* & 0 & \cdots & 0 \\ 0 & l_2^* = s_2^* q_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_r^* = s_r^* q_r^* \end{pmatrix} \quad (3.29)$$

From equation 3.28, suppose $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_r^{(t)})$ at current t^{th} iteration and $\Theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_r^*)$, each $s_i^* = 1 - m_i^*$ is really the speed that the next $\theta_i^{(t+1)}$ moves to θ_i^*. From equation 3.26 and equation 3.28, equation 3.30 specifies the speed s^* of GEM algorithm.

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

Where,

$$m^* = \max_{m_i^*} \{m_1^*, m_2^*, \ldots, m_r^*\} \quad (3.30)$$

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 Θ degrades into scalar as $\Theta = \theta$, the fourth column of table 1.2 (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 3.23 with $\theta^* = 0.6268$.

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

In practice, if GEM is run step by step, θ^* is not known yet at some t^{th} iteration when GEM does not converge yet. Hence, equation 3.24 (McLachlan & Krishnan, 1997, p. 120) is used to make approximation of $M_e^* = DM(\theta^*)$ with unknown θ^* and $\theta^{(t)} \neq \theta^{(t+1)}$.

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

It is required only two successive iterations because both $\theta^{(t)}$ and $\theta^{(t+1)}$ are determined at t^{th} iteration whereas $\theta^{(t+2)}$ is determined at $(t+1)^{\text{th}}$ iteration. For example, in table 1.2, given $\theta^{(1)} = 0.5$, $\theta^{(2)} = 0.6082$, and $\theta^{(3)} = 0.6243$, at $t = 1$, we have:

$$DM(\theta^*) \approx \frac{\theta^{(3)} - \theta^{(2)}}{\theta^{(2)} - \theta^{(1)}} = \frac{0.6243 - 0.6082}{0.6082 - 0.5} = 0.1488$$

Whereas the real $M_e^* = DM(\theta^*)$ is 0.1465 shown in the fourth column of table 1.2 at $t = 1$.

We will prove by contradiction that if definition 3.1 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 $Q(M(\Theta^{(t)}) \mid \Theta^{(t)}) > Q(\Theta^{(t)} \mid \Theta^{(t)})$. Given Θ degrades into scalar as $\Theta = \theta$ with note that scalar is 1-element vector, 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 3.29, we deduce that $D^2L(\theta^*) = L_e^* = S_e^* Q_e^* < 0$ where $Q_e^* = D^{20}Q(\theta^* \mid \theta^*) < 0$. However, this contradicts the converse assumption "there always exists $l_i^* \geq 0$ for some i when $Q(M(\Theta^{(t)}) \mid \Theta^{(t)}) > Q(\Theta^{(t)} \mid \Theta^{(t)})$". Therefore, if $Q(M(\Theta^{(t)}) \mid \Theta^{(t)}) > Q(\Theta^{(t)} \mid \Theta^{(t)})$ then, $l_i^* < 0$ for all

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

i. In other words, at that time, $D^2L(\Theta^*) = L_e^*$ is negative definite. Recall that we proved that $DL(\Theta^*) = 0$ for corollary 3.3. Now we have $D^2L(\Theta^*)$ negative definite, which means that Θ^* is a local maximizer of $L(\Theta^*)$ in corollary 3.3. In other words, corollary 3.3 is proved.

Recall that $L(\Theta)$ is the log-likelihood function of observed Y according to equation 2.3.

$$L(\Theta) = \log\big(g(Y|\Theta)\big) = \log\left(\int_{\varphi^{-1}(Y)} f(X|\Theta)\mathrm{d}X\right)$$

Both $-D^{20}H(\Theta^* \mid \Theta^*)$ and $-D^{20}Q(\Theta^* \mid \Theta^*)$ are information matrices (Zivot, 2009, pp. 7-9) specified by equation 3.31.

$$\begin{aligned} I_H(\Theta^*) &= -D^{20}H(\Theta^*|\Theta^*) \\ I_Q(\Theta^*) &= -D^{20}Q(\Theta^*|\Theta^*) \end{aligned} \tag{3.31}$$

$I_H(\Theta^*)$ measures information of X about Θ^* with support of Y whereas $I_Q(\Theta^*)$ measures information of X about Θ^*. In other words, $I_H(\Theta^*)$ measures observed information whereas $I_Q(\Theta^*)$ measures hidden information. Let $V_H(\Theta^*)$ and $V_Q(\Theta^*)$ be covariance matrices of Θ^* with regard to $I_H(\Theta^*)$ and $I_Q(\Theta^*)$, respectively. They are inverses of $I_H(\Theta^*)$ and $I_Q(\Theta^*)$ according to equation 3.32 when Θ^* is unbiased estimate.

$$\begin{aligned} V_H(\Theta^*) &= \big(I_H(\Theta^*)\big)^{-1} \\ V_Q(\Theta^*) &= \big(I_Q(\Theta^*)\big)^{-1} \end{aligned} \tag{3.32}$$

Equation 3.33 is a variant of equation 3.17 to calculate $DM(\Theta^*)$ based on information matrices:

$$DM(\Theta^*) = I_H(\Theta^*)\big(I_Q(\Theta^*)\big)^{-1} = \big(V_H(\Theta^*)\big)^{-1}V_Q(\Theta^*) \tag{3.33}$$

If $f(X \mid \Theta)$, $g(Y \mid \Theta)$ and $k(X \mid Y, \Theta)$ belong to exponential family, from equation 3.14 and equation 3.16, we have:

$$\begin{aligned} D^{20}H(\Theta^*|\Theta^*) &= -V(\tau(X)|Y,\Theta^*) \\ D^{20}Q(\Theta^*|\Theta^*) &= -V(\tau(X)|\Theta^*) \end{aligned}$$

Hence, equation 3.34 specifies $DM(\Theta^*)$ in case of exponential family.

$$DM(\Theta^*) = V(\tau(X)|Y,\Theta^*)\big(V(\tau(X)|\Theta^*)\big)^{-1} \tag{3.34}$$

Equation 3.35 specifies relationships among $V_H(\Theta^*)$, $V_Q(\Theta^*)$, $V(\tau(X) \mid Y, \Theta^*)$, and $V(\tau(X) \mid \Theta^*)$ in case of exponential family.

$$\begin{aligned} V_H(\Theta^*) &= \big(V(\tau(X)|Y,\Theta^*)\big)^{-1} \\ V_Q(\Theta^*) &= \big(V(\tau(X)|\Theta^*)\big)^{-1} \end{aligned} \tag{3.35}$$

4. Variants of EM algorithm

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

4.1. EM algorithm with prior probability

DLR (Dempster, Laird, & Rubin, 1977, pp. 6, 11) mentioned that the convergence rate $DM(\Theta^*)$ specified by equation 3.17 can be improved by adding a prior probability $\pi(\Theta)$ in conjugation with $f(X \mid \Theta)$, $g(Y \mid \Theta)$ or $k(X \mid Y, \Theta)$ according to maximum a posteriori probability (MAP) method (Wikipedia, Maximum a posteriori estimation, 2017). For example, if $\pi(\Theta)$ in conjugation with $g(Y \mid \Theta)$ then, the posterior probability $\pi(\Theta \mid Y)$ is:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

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

It implies from equation 3.2 that

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

Let,

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

GEM algorithm now aims to maximize $Q_+(\Theta' \mid \Theta)$ instead of maximizing $Q(\Theta' \mid \Theta)$. The proof of convergence for $Q_+(\Theta' \mid \Theta)$ is not changed in manner but determining the convergence matrix M_e for $Q_+(\Theta' \mid \Theta)$ is necessary. Because $H(\Theta' \mid \Theta)$ is kept intact whereas $Q(\Theta' \mid \Theta)$ is replaced by $Q_+(\Theta' \mid \Theta)$, we expect that the convergence rate m^* specified by equation 3.26 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 an (orthogonal) eigenvector matrix V such that (Wikipedia, Diagonalizable matrix, 2017) (StackExchange, 2013):

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

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

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

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

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

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

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

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

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

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 π_i^* are non-positive, we have:

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

Therefore, by comparing equation 4.1.6 and equation 3.26, we conclude that m^* is smaller with regard to $Q_+(\Theta' \mid \Theta)$. In other words, the convergence rate is improved with support of prior probability $\pi(\Theta)$. In literature of EM, the combination of GEM and MAP with support of $\pi(\Theta)$ results out a so-called MAP-GEM algorithm.

4.2. EM algorithm with Newton-Raphson method

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

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

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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 implies:

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

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 Θ_{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.

E-step:

The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$, according to equation 2.8. Actually, $Q(\Theta \mid \Theta^{(t)})$ is formulated as function of Θ.

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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 an (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 $Q_e{}^{(t+1)}$. Suppose $D^{20}Q(\Theta_0{}^{(t+1)} \mid \Theta^{(t)})$ 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:

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

Hence, EM1 surely converges to a local maximizer Θ^* according to corollary 3.3 with assumption that $D^{20}Q(\Theta_0{}^{(t+1)} \mid \Theta^{(t)})$ and $D^{20}Q(\Theta^{(t)} \mid \Theta^{(t)})$ are negative definite for all t where $\Theta_0{}^{(t+1)}$ is a point on the line segment joining Θ and $\Theta^{(t+1)}$.

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 2.3) and EM1 (table 4.2.1), we conclude that EM1 increases $Q(\Theta \mid \Theta^{(t)})$ after each iteration whereas GEM maximizes $Q(\Theta \mid \Theta^{(t)})$ after each iteration. However, EM1 will maximizes $Q(\Theta \mid \Theta^{(t)})$ at the last iteration when it converges. EM1 gains this excellent and interesting result because of Newton-Raphson process specified by equation 4.2.2.

Because equation 3.17 is 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: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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 $D^{20}Q(\Theta^{(1)} \mid \Theta^{(1)})$ over all iterations where $\Theta^{(1)}$ 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).

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

In equation 4.2.3, only $D^{10}Q(\Theta^{(t)} \mid \Theta^{(t)})$ is re-computed at every iteration whereas $D^{20}Q(\Theta^{(1)} \mid \Theta^{(1)})$ is fixed. Equation 4.2.3 implies a pseudo Newton-Raphson process which still converges to a local maximizer Θ^{*} but it is slower than Newton-Raphson process specified by equation 4.2.2 (Ta, 2014).

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

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

Where, (4.2.4)

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

The convergence of equation 4.2.4 is same as the convergence of equation 4.2.2. 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:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

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

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:

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

This means:

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

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

E-step:

The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$, according to equation 2.8. Actually, $Q(\Theta \mid \Theta^{(t)})$ is formulated as function of Θ.

M-step:

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

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

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

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

Where,

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

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:

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

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

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

As a result, we have:

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

Hence, GD-GEM surely converges to a local maximizer Θ^* according to corollary 3.3 with assumption that $D^{20}Q(\Theta_0^{(t+1)} \mid \Theta^{(t)})$ is negative definite where $\Theta_0^{(t+1)}$ is a point on the line segment joining Θ and $\Theta^{(t+1)}$.

It is not easy to solve the maximization problem with regard to γ 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 Θ is scalar such that $\Theta = \theta$, the sequence $\{\theta^{(t)}\}_{t=1}^{+\infty} = \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}, \ldots$ is monotonous. From equation 3.23

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

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):

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

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

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

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

Where Δ is forward difference operator,

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

And

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

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

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

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

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

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 2.8. Actually, $Q(\Theta \mid \Theta^{(t)})$ is formulated as function of Θ. Note that $t = 1, 2, 3, \ldots$ and $\Theta^{(0)} = \Theta^{(1)}$.

M-step:

Let $\Theta^{(t+1)} = (\theta_1^{(t+1)}, \theta_2^{(t)}, \ldots, \theta_r^{(t+1)})^T$ be a maximizer of $Q(\Theta \mid \Theta^{(t)})$. Note $\Theta^{(t+1)}$ will become current parameter at the next iteration ($(t+1)^{\text{th}}$ iteration).

Aitken parameter $\widehat{\Theta}^{(t-1)} = \left(\hat{\theta}_1^{(t-1)}, \hat{\theta}_2^{(t-1)}, \ldots, \hat{\theta}_r^{(t-1)}\right)^T$ is calculated according to equation 4.3.2.

$$\hat{\theta}_i^{(t-1)} = \theta_i^{(t-1)} - \frac{\left(\Delta\theta_i^{(t-1)}\right)^2}{\Delta^2\theta_i^{(t-1)}}$$

If $\widehat{\Theta}^{(t-1)} = \widehat{\Theta}^{(t-2)}$ then, the algorithm stops and we have $\widehat{\Theta}^{(t-1)} = \widehat{\Theta}^{(t-2)} = \Theta^*$.

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

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

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

4.4. ECM algorithm

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

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

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

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

E-step:

As usual, $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$ according to equation 2.8. Actually, $Q(\Theta \mid \Theta^{(t)})$ is formulated as function of Θ.

CM-steps:

There are S CM-steps. In every s^{th} CM step (s =1, 2,…, S), finding

$$\Theta^{\left(t+\frac{s}{S}\right)} = \underset{\Theta}{\operatorname{argmax}}\left\{Q\left(\Theta\middle|\Theta^{(t)}\right) \text{ with subject to } g_s(\Theta) = g_s\left(\Theta^{\left(t+\frac{s-1}{S}\right)}\right)\right\} \tag{4.4.2}$$

The next parameter $\Theta^{(t+1)}$ is the output of the final CM-step (S^{th} CM-step):

$$\Theta^{(t+1)} = \Theta^{\left(t+\frac{S}{S}\right)} \tag{4.4.3}$$

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

From equation 4.4.2 and equation 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 3.3. However, Meng and Rubin (Meng & Rubin, 1993, pp. 274-276) provided some conditions for convergence of ECM to a maximizer of $L(\Theta)$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

5. Applications of EM

5.1. Mixture model and EM

As usual, let $\boldsymbol{X}$ be the hidden or latent space and let $\boldsymbol{Y}$ be the observed space. Especially, the random variable X in $\boldsymbol{X}$ represents latent class or latent component of random variable Y in $\boldsymbol{Y}$. Suppose X is discrete and ranges in $\boldsymbol{X} = \{1, 2, \ldots, K\}$. The so-called probabilistic finite *mixture model* is represented by the PDF of Y, as seen in equation 5.1.1.

$$f(Y|\Theta) = \sum_{X=1}^{K} \alpha_X f_X(Y|\theta_X) \tag{5.1.1}$$

Where,

$$\Theta = (\alpha_1, \alpha_2, \ldots, \alpha_K, \theta_1, \theta_2, \ldots, \theta_K)^T$$

$$\sum_{k=1}^{K} \alpha_k = 1$$

Note, Y can be discrete or continuous. Recall that the ultimate purpose of EM algorithm is to maximize $f(Y|\Theta)$ with subject to Θ. Each $f_X(Y|\theta_X)$ is called the X^{th} partial PDF of Y whose partial parameter is θ_X. Each $f_X(Y|\theta_X)$ is also called the X^{th} observational PDF of Y. It is really the conditional PDF of Y given X, as seen in equation 5.1.2.

$$f_X(Y|\theta_X) = f(Y|X, \theta_X) \tag{5.1.2}$$

From equation 5.1.1, the mixture model $f(Y|\Theta)$ is the mean of K partial PDFs. The variable X implies which partial PDF "generates" Y (Bilmes, 1998, p. 5).

Each α_X is called mixture coefficient. It is really the probability of discrete X, as seen in equation 5.1.3. However, in mixture model, each α_X is also considered as parameter, which is belongs to the compound parameter Θ.

$$\alpha_X = P(X) \tag{5.1.3}$$

The joint probabilistic distribution of X and Y, which implies the implicit mapping between $\boldsymbol{X}$ and $\boldsymbol{Y}$, is product of the mixture coefficient α_X and the X^{th} PDF of Y, as seen in equation 5.1.4.

$$f(X, Y|\Theta) = P(X) f(Y|X, \theta_X) = \alpha_X f_X(Y|\theta_X) \tag{5.1.4}$$

This implies:

$$f(Y|\Theta) = \sum_{X=1}^{K} f(X, Y|\Theta) = \sum_{X=1}^{K} P(X) f(Y|X, \theta_X) = \sum_{X=1}^{K} \alpha_X f_X(Y|\theta_X) \tag{5.1.5}$$

Equation 5.1.6 specifies the conditional probability of X given Y. Please pay attention to this important probability.

$$P(X|Y, \Theta) = \frac{\alpha_X f_X(Y|\theta_X)}{\sum_{l=1}^{K} \alpha_l f_l(Y|\theta_l)} \tag{5.1.6}$$

Following is the proof of equation 5.1.6. According to Bayes' rule, we have:

$$P(X = x|Y = y, \Theta) = \frac{P(x) f(y|x, \theta_x)}{\sum_{X=1}^{K} P(X) f(Y|X, \theta_X)}$$

Applying equation 5.1.3 and equation 5.1.4, we have:

$$P(X = x|Y = y, \Theta) = \frac{\alpha_x f_x(y|\theta_x)}{\sum_{X=1}^{K} \alpha_X f_X(Y|\theta_X)}$$

In other words, equation 5.1.6 is established■

Now GEM algorithm is applied into mixture model for estimating the parameter Θ. Derived from equation 2.12, the conditional expectation $Q(\Theta'|\Theta)$ of mixture model becomes:

$$Q(\Theta'|\Theta) = \sum_{X \in \boldsymbol{X}} P(X|Y, \Theta) \log\big(f(X, Y|\Theta')\big) = \sum_{X \in \boldsymbol{X}} P(X|Y, \Theta) \log\big(\alpha_X f_X(Y|\theta'_X)\big) \tag{5.1.7}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

In practice, suppose Y is observed as a sample $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$ of size N with note that all Y_i (s) are mutually independent and identically distributed (iid). The observed sample $\mathcal{Y}$ is associated with a a hidden set (latent set) $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ of size N. All X_i (s) are iid and they are not existent in fact. Let $X \in \boldsymbol{X}$ be the random variable representing every X_i. Of course, the domain of X is $\boldsymbol{X}$. Derived from equation 2.15, equation 5.1.8 specifies $Q(\Theta'|\Theta)$ given such $\mathcal{Y}$.

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N} \sum_{X \in \boldsymbol{X}} P(X|Y_i, \Theta) \log\left(\alpha_X f_X(Y_i|\theta'_X)\right) \tag{5.1.8}$$

Equation 5.1.8 is the general case of equation 5.1.7. At the t^{th} iteration of GEM, given current parameter $\Theta^{(t)} = (\alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_K^{(t)}, \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_K^{(t)})^T$, the conditional expectation specified by equation 5.1.8 is written as follows:

$$Q\left(\Theta\middle|\Theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{X \in \boldsymbol{X}} P\left(X\middle|Y_i, \Theta^{(t)}\right) \log\left(\alpha_X f_X(Y_i|\theta_X)\right)$$

Thus, the unknown of $Q(\Theta|\Theta^{(t)})$ is $\Theta = (\alpha_1, \alpha_2, \ldots, \alpha_K, \theta_1, \theta_2, \ldots, \theta_K)^T$. Because X is discrete and ranges in $\{1, 2, \ldots, K\}$, the conditional expectation $Q(\Theta|\Theta^{(t)})$ is re-written as equation 5.1.9 for convenience.

$$Q\left(\Theta\middle|\Theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} P\left(k\middle|Y_i, \Theta^{(t)}\right) \log\left(\alpha_k f_k(Y_i|\theta_k)\right) \tag{5.1.9}$$

Where the conditional probability $P(k \mid Y, \Theta^{(t)})$ is determined by equation 5.1.10 which is indeed equation 5.1.6.

$$P\left(k\middle|Y_i, \Theta^{(t)}\right) = P\left(X = k\middle|Y_i, \Theta^{(t)}\right) = \frac{\alpha_k^{(t)} f_k\left(Y_i\middle|\theta_k^{(t)}\right)}{\sum_{l=1}^{K} \alpha_l^{(t)} f_l\left(Y_i\middle|\theta_l^{(t)}\right)} \tag{5.1.10}$$

At M-step of the current t^{th} iteration, $Q(\Theta|\Theta^{(t)})$ specified by equation 5.1.9 is maximized with subject to Θ. How to maximize $Q(\Theta|\Theta^{(t)})$ with subject to Θ is dependent on types of partial PDFs $f_k(Y_i|\theta_k)$.

Because there is the constraint $\sum_{k=1}^{n} \theta_k = 1$, we use Lagrange duality method to maximize to maximize $Q(\Theta|\Theta^{(t)})$. The Lagrange function $la(\Theta, \lambda \mid \Theta^{(t)})$ is sum of $Q(\Theta|\Theta^{(t)})$ and the constraint $\sum_{k=1}^{K} \alpha_k = 1$, which is specified by equation 5.1.11.

$$\begin{aligned} la\left(\Theta, \lambda\middle|\Theta^{(t)}\right) &= Q\left(\Theta\middle|\Theta^{(t)}\right) + \lambda\left(1 - \sum_{k=1}^{K} \alpha_k\right) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} P\left(k\middle|Y_i, \Theta^{(t)}\right) \log(\alpha_k) \\ &+ \sum_{i=1}^{N} \sum_{k=1}^{K} P\left(k\middle|Y_i, \Theta^{(t)}\right) \log\left(f_k(Y_i|\theta_k)\right) + \lambda\left(1 - \sum_{k=1}^{K} \alpha_k\right) \end{aligned} \tag{5.1.11}$$

Note, $\lambda \geq 0$ is called Lagrange multiplier. Of course, $la(\Theta, \lambda \mid \Theta^{(t)})$ is function of Θ and λ. The next parameters $\alpha_k^{(t+1)}$ that maximizes $Q(\Theta|\Theta^{(t)})$ is solution of the equation formed by setting the first-order partial derivative of Lagrange function regarding α_k and λ to be zero with suppose that the Lagrange function is first-order smooth function.

$$\frac{\partial la\left(\Theta, \lambda\middle|\Theta^{(t)}\right)}{\partial \alpha_k} = 0$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\Leftrightarrow \frac{\partial}{\partial \alpha_k}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} P\left(k\middle|Y_i,\Theta^{(t)}\right)\log(\alpha_k)+\lambda\left(1-\sum_{k=1}^{K}\alpha_k\right)\right)=0$$

$$\Leftrightarrow \sum_{i=1}^{N}\frac{1}{\alpha_k}P\left(k\middle|Y_i,\Theta^{(t)}\right)-\lambda=0$$

This implies:

$$\sum_{i=1}^{N} P\left(k\middle|Y_i,\Theta^{(t)}\right)-\alpha_k\lambda=0 \tag{5.1.12}$$

Summing equation 5.1.12 over K classes $\{1, 2,\ldots, K\}$, we have (Bilmes, 1998, p. 5):

$$\sum_{i=1}^{N}\sum_{k=1}^{K} P\left(k\middle|Y_i,\Theta^{(t)}\right)-\lambda\sum_{k=1}^{K}\alpha_k=0$$

$$\Leftrightarrow N-\lambda=0$$

$$\left(\text{due to } \sum_{k=1}^{K} P\left(k\middle|Y_i,\Theta^{(t)}\right)=1 \text{ and } \sum_{k=1}^{K}\alpha_k=1\right)$$

$$\Leftrightarrow \lambda=N$$

Substituting $\lambda = N$ into equation 5.1.12, the next parameters $\alpha_k^{(t+1)}$ is totally determined by equation 5.1.13.

$$\alpha_k^{(t+1)}=\frac{1}{N}\sum_{i=1}^{N} P\left(k\middle|Y_i,\Theta^{(t)}\right) \tag{5.1.13}$$

Note, the conditional probability $P(k \mid Y_i, \Theta^{(t)})$ is determined by equation 5.1.10.

When parameters $\alpha_k^{(t+1)}$ and λ are determined, the Lagrange function $la(\Theta, \lambda \mid \Theta^{(t)})$ is now function of parameters θ_k as $la(\theta_k|\theta_k^{(t)})$. The next parameters $\theta_k^{(t+1)}$ is solution of the equation formed by setting the first-order partial derivative of Lagrange function regarding θ_k to be zero with suppose that the Lagrange function is first-order smooth function.

$$\frac{\partial la\left(\Theta,\lambda\middle|\Theta^{(t)}\right)}{\partial\theta_k}=0$$

$$\Leftrightarrow \frac{\partial}{\partial\theta_k}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} P\left(k\middle|Y_i,\Theta^{(t)}\right)\log\left(f_k(Y_i|\theta_k)\right)\right)=0$$

$$\Leftrightarrow \sum_{i=1}^{N} P\left(k\middle|Y_i,\Theta^{(t)}\right)\frac{\partial\log\left(f_k(Y_i|\theta_k)\right)}{\partial\theta_k}=0$$

Thus, the next parameters $\theta_k^{(t+1)}$ is solution of the equation 5.1.14.

$$\sum_{i=1}^{N} P\left(k\middle|Y_i,\Theta^{(t)}\right)\frac{\partial\log\left(f_k(Y_i|\theta_k)\right)}{\partial\theta_k}=0 \tag{5.1.14}$$

The two steps of GEM algorithm for constructing mixture model at some t^{th} iteration are shown in table 5.1.1. Note, suppose the Lagrange function is first-order smooth function.

E-step:

The conditional probability $P(k \mid Y_i, \Theta^{(t)})$ is calculated based on current parameter $\Theta^{(t)} = (\alpha_1^{(t)}, \alpha_2^{(t)},\ldots, \alpha_K^{(t)}, \theta_1^{(t)}, \theta_2^{(t)},\ldots, \theta_K^{(t)})^T$, according to equation 5.1.10.

$$P\left(k\middle|Y_i,\Theta^{(t)}\right)=\frac{\alpha_k^{(t)}f_k\left(Y_i\middle|\theta_k^{(t)}\right)}{\sum_{l=1}^{K}\alpha_l^{(t)}f_l\left(Y_i\middle|\theta_l^{(t)}\right)}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

M-step:

The next parameter $\Theta^{(t+1)} = (\alpha_1^{(t+1)}, \alpha_2^{(t+1)},\ldots, \alpha_K^{(t+1)}, \theta_1^{(t+1)}, \theta_2^{(t+1)},\ldots, \theta_K^{(t+1)})^T$, which is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with subject to Θ, is calculated by equation 5.1.13 and equation 5.1.14. Note, $\theta_k^{(t+1)}$ is solution of the equation 5.1.14.

$$\alpha_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right)$$

$$\theta_k^{(t+1)}: \sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right) \frac{\partial \log\left(f_k\left(Y_i\middle|\theta_k^{(t+1)}\right)\right)}{\partial \theta_k} = 0$$

Table 5.1.1. E-step and M-step of GEM algorithm for constructing mixture model regarding first-order smooth Lagrange function

GEM algorithm converges at some t^{th} iteration. At that time, $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ is the optimal estimate of mixture model regarding first-order smooth Lagrange function.

Suppose that each PDF $f_k(Y_i|\theta_k)$) belongs to regular exponential family and then, solving equation 5.1.4 is easier as follows:

$$\sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right) \frac{\partial \log\left(f_k(Y_i|\theta_k)\right)}{\partial \theta_k} = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right) \frac{\partial \log\left(b(Y_i)\exp\left(\theta_k^T \tau(Y_i)\right)/a(\theta_k)\right)}{\partial \theta_k} = 0$$

(Due to $f_k(Y_i|\theta_k)$) belongs to exponential family)

$$\Leftrightarrow \sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right)\left(\tau(Y_i) - \log'\left(a(\theta_k)\right)\right) = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right)\left(\tau(Y_i) - E(\tau(Y)|\theta_k)\right) = 0$$

(Due to $\log'(a(\theta_k)) = E(\tau(Y|\theta_k))$, please see table 1.2)

In general, the next parameters $\theta_k^{(t+1)}$ is solution of the equation 5.1.15 within regular exponential family.

$$\sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right)\left(\tau(Y_i) - E(\tau(Y)|\theta_k)\right) = 0 \quad (5.1.15)$$

Where Y is the random variable representing all Y_i (s) and,

$$E(\tau(Y)|\theta_k) = \int_Y \tau(Y) f_k(Y|\theta_k) \mathrm{d}Y$$

The two steps of GEM algorithm for constructing mixture model at some t^{th} iteration are shown in table 5.1.2 with suppose that each partial PDF $f_X(Y|\theta_X)$ is assumed to belong regular exponential family.

E-step:

The conditional probability $P(k \mid Y_i, \Theta^{(t)})$ is calculated based on current parameter $\Theta^{(t)} = (\alpha_1^{(t)}, \alpha_2^{(t)},\ldots, \alpha_K^{(t)}, \theta_1^{(t)}, \theta_2^{(t)},\ldots, \theta_K^{(t)})^T$, according to equation 5.1.10.

$$P\left(k\middle|Y_i, \Theta^{(t)}\right) = \frac{\alpha_k^{(t)} f_k\left(Y_i\middle|\theta_k^{(t)}\right)}{\sum_{l=1}^{K} \alpha_l^{(t)} f_l\left(Y_i\middle|\theta_l^{(t)}\right)}$$

M-step:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

The next parameter $\Theta^{(t+1)} = (\alpha_1^{(t+1)}, \alpha_2^{(t+1)}, \ldots, \alpha_K^{(t+1)}, \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_K^{(t+1)})^T$, which is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with subject to Θ, is calculated by equation 5.1.13 and equation 5.1.15. Note, $\theta_k^{(t+1)}$ is solution of the equation 5.1.15.

$$\alpha_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} P\left(k \middle| Y_i, \Theta^{(t)}\right)$$

$$\theta_k^{(t+1)}: \sum_{i=1}^{N} P\left(k \middle| Y_i, \Theta^{(t)}\right)\left(\tau(Y_i) - E\left(\tau(Y)\middle|\theta_k^{(t+1)}\right)\right) = 0$$

Table 5.1.1. E-step and M-step of GEM algorithm for constructing mixture model regarding regular exponential family

GEM algorithm converges at some t^{th} iteration. At that time, $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ is the optimal estimate of mixture model regarding regular exponential family.

There is a special case that each $f_k(Y_i|\theta_k)$ is normal distribution, which is popular in domain of mixture model, with note that normal distribution belongs to regular exponential family. Thus, let Y be random variable representing all Y_i. Without loss of generality, suppose Y is vector so that each $f_k(Y|\theta_k)$ is multivariate normal distribution. Recall that each $f_k(Y|\theta_k)$ is called the k^{th} partial PDF of Y or the k^{th} observational PDF of Y. In this case, the mixture model is called *normal mixture model* (Gaussian mixture model) and it is easy to solve equation 5.1.14 or equation 5.1.15 for θ_k. Suppose random variable Y is vector of size n.

$$f_k(Y|\theta_k) = (2\pi)^{-\frac{n}{2}}|\Sigma_k|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(Y-\mu_k)^T\Sigma_k^{-1}(Y-\mu_k)\right) \quad (5.1.16)$$

Where μ_k and Σ_k are mean vector and covariance matrix of $f_k(Y|\theta_k)$, respectively. The notation $|.|$ denotes determinant of given matrix and the notation Σ_k^{-1} denotes inverse of matrix Σ_k. Note, Σ_k is invertible and symmetric. Now we find other parameters $\theta_k^{(t+1)} = (\mu_k^{(t+1)}, \Sigma_k^{(t+1)})^T$ by solving directly equation 5.1.14 or equation 5.1.15. Recall that each Y_i conforms to multivariate normal distribution, according to equation 5.1.16.

$$f_k(Y_i|\theta_k) = (2\pi)^{-\frac{n}{2}}|\Sigma_k|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(Y_i-\mu_k)^T\Sigma_k^{-1}(Y_i-\mu_k)\right)$$

Where μ_k and Σ_k are mean and covariance matrix of $f_k(Y_i|\theta_k)$, respectively. The Lagrange function is re-written as follows:

$$la\left(\Theta, \lambda \middle| \Theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} P\left(k \middle| Y_i, \Theta^{(t)}\right)\log(\alpha_k) + \sum_{i=1}^{N}\sum_{k=1}^{K} P\left(k \middle| Y_i, \Theta^{(t)}\right)\left(-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(Y_i-\mu_k)^T\Sigma_k^{-1}(Y_i-\mu_k)\right) + \lambda\left(1 - \sum_{k=1}^{K}\alpha_k\right)$$

Where p is the dimension of Y_i; in other words, p is the dimension of space $\boldsymbol{Y}$.

The first-order partial derivative of Lagrange function with respect to μ_k is (Nguyen, 2015, p. 35):

$$\frac{\partial la\left(\Theta, \lambda \middle| \Theta^{(t)}\right)}{\partial \mu_k} = \sum_{i=1}^{N} P\left(k \middle| Y_i, \Theta^{(t)}\right)((Y_i-\mu_k)^T\Sigma_k^{-1})$$

$$\left(\text{due to } \frac{\partial (Y_i-\mu_k)^T\Sigma_k^{-1}(Y_i-\mu_k)}{\partial \mu_k} = -2(Y_i-\mu_k)^T\Sigma_k^{-1} \text{ when } \Sigma_k^{-1} \text{ is symmetric}\right)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

The next parameter $\mu_k^{(t+1)}$ that maximizes $Q(\Theta|\Theta^{(t)})$ is solution of the equation formed by setting the first-order partial derivative of Lagrange function with regard to μ_k to be $\mathbf{0}^T$. Note that $\mathbf{0} = (0, 0,\ldots, 0)^T$ is zero vector.

$$\frac{\partial la(\Theta,\lambda|\Theta^{(t)})}{\partial \mu_k} = \mathbf{0}^T$$

$$\Leftrightarrow \sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})((Y_i-\mu_k)^T\Sigma_k^{-1}) = \mathbf{0}^T$$

$$\Leftrightarrow \left(\sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})(Y_i-\mu_k)^T\right)\Sigma_k^{-1} = \mathbf{0}^T$$

$$\Rightarrow \sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})(Y_i-\mu_k)^T = \mathbf{0}^T$$

$$\Leftrightarrow \sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})(Y_i-\mu_k) = \mathbf{0}$$

$$\Leftrightarrow \sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})Y_i - \left(\sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})\right)\mu_k = \mathbf{0}$$

$$\Leftrightarrow \left(\sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})\right)\mu_k = \sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})Y_i$$

This implies equation 5.1.17 to specify the next parameter $\mu_k^{(t+1)}$.

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})Y_i}{\sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})} \quad (5.1.17)$$

Note, the conditional probability $P(k \mid Y_i, \Theta^{(t)})$ is determined by equation 5.1.10.

The first-order partial derivative of Lagrange function with respect to Σ_k is:

$$\frac{\partial la(\Theta,\lambda|\Theta^{(t)})}{\partial \Sigma_k} = \sum_{i=1}^{N} P(k|Y_i,\Theta^{(t)})\left(-\frac{1}{2}\Sigma_k^{-1} + \frac{1}{2}\Sigma_k^{-1}(Y_i-\mu_k)(Y_i-\mu_k)^T\Sigma_k^{-1}\right)$$

Due to:

$$\frac{\partial \log(|\Sigma_k|)}{\partial \Sigma_k} = \Sigma_k^{-1}$$

And

$$\frac{\partial (Y_i-\mu_k)^T\Sigma_k^{-1}(Y_i-\mu_k)}{\partial \Sigma_k} = \frac{\partial \mathrm{tr}\left((Y_i-\mu_k)(Y_i-\mu_k)^T\Sigma_k^{-1}\right)}{\partial \Sigma_k}$$

Because Bilmes (Bilmes, 1998, p. 5) mentioned:

$$(Y_i-\mu_k)^T\Sigma_k^{-1}(Y_i-\mu_k) = \mathrm{tr}\left((Y_i-\mu_k)(Y_i-\mu_k)^T\Sigma_k^{-1}\right)$$

Where tr(A) is trace operator which takes sum of diagonal elements of matrix $\mathrm{tr}(A) = \sum_i a_{ii}$.

This implies (Nguyen, 2015, p. 45):

$$\frac{\partial (Y_i-\mu_k)^T\Sigma_k^{-1}(Y_i-\mu_k)}{\partial \Sigma_k} = \frac{\partial \mathrm{tr}\left((Y_i-\mu_k)(Y_i-\mu_k)^T\Sigma_k^{-1}\right)}{\partial \Sigma_k} = -\Sigma_k^{-1}(Y_i-\mu_k)(Y_i-\mu_k)^T\Sigma_k^{-1}$$

Where Σ_k is symmetric and invertible matrix. Substituting the next parameter $\mu_k^{(t+1)}$ specified by equation 5.1.16 into the first-order partial derivative of Lagrange function with respect to Σ_k, we have:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\frac{\partial la\left(\Theta,\lambda\middle|\Theta^{(t)}\right)}{\partial\Sigma_k}=\sum_{i=1}^{N}P\left(k\middle|Y_i,\Theta^{(t)}\right)\left(-\frac{1}{2}\Sigma_k^{-1}+\frac{1}{2}\Sigma_k^{-1}\left(Y_i-\mu_k^{(t+1)}\right)\left(Y_i-\mu_k^{(t+1)}\right)^T\Sigma_k^{-1}\right)$$

The next parameter $\Sigma_k^{(t+1)}$ that maximizes $Q(\Theta|\Theta^{(t)})$ is the solution of equation formed by setting the first-order partial derivative of Lagrange function regarding Σ_k to zero matrix. Let $(\mathbf{0})$ denote zero matrix.

$$(\mathbf{0})=\begin{pmatrix}0&0&\cdots&0\\0&0&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&0\end{pmatrix}$$

We have:

$$\frac{\partial la\left(\Theta,\lambda\middle|\Theta^{(t)}\right)}{\partial\Sigma_k}=(\mathbf{0})$$

$$\Leftrightarrow\sum_{i=1}^{N}P\left(k\middle|Y_i,\Theta^{(t)}\right)\left(-\frac{1}{2}\Sigma_k^{-1}+\frac{1}{2}\Sigma_k^{-1}\left(Y_i-\mu_k^{(t+1)}\right)\left(Y_i-\mu_k^{(t+1)}\right)^T\Sigma_k^{-1}\right)=(\mathbf{0})$$

$$\Rightarrow\sum_{i=1}^{N}P\left(k\middle|Y_i,\Theta^{(t)}\right)\left(-\Sigma_k+\left(Y_i-\mu_k^{(t+1)}\right)\left(Y_i-\mu_k^{(t+1)}\right)^T\right)=(\mathbf{0})$$

$$\Leftrightarrow\sum_{i=1}^{N}P\left(k\middle|Y_i,\Theta^{(t)}\right)\left(\left(Y_i-\mu_k^{(t+1)}\right)\left(Y_i-\mu_k^{(t+1)}\right)^T\right)-\left(\sum_{i=1}^{N}P\left(k\middle|Y_i,\Psi^{(t)}\right)\right)\Sigma_k=(\mathbf{0})$$

$$\Leftrightarrow\left(\sum_{i=1}^{N}P\left(k\middle|Y_i,\Theta^{(t)}\right)\right)\Sigma_k=\sum_{i=1}^{N}P\left(k\middle|Y_i,\Theta^{(t)}\right)\left(\left(Y_i-\mu_k^{(t+1)}\right)\left(Y_i-\mu_k^{(t+1)}\right)^T\right)$$

This implies equation 5.1.18 to specify the next parameter $\Sigma_k^{(t+1)}$.

$$\Sigma_k^{(t+1)}=\frac{\sum_{i=1}^{N}P\left(k\middle|Y_i,\Theta^{(t)}\right)\left(\left(Y_i-\mu_k^{(t+1)}\right)\left(Y_i-\mu_k^{(t+1)}\right)^T\right)}{\sum_{i=1}^{N}P\left(k\middle|Y_i,\Theta^{(t)}\right)}\qquad(5.1.18)$$

Note, the conditional probability $P(k \mid Y_i, \Theta^{(t)})$ is determined by equation 5.1.10 and the next parameter $\mu_k^{(t+1)}$ is specified by equation 5.1.17.

As a result, the solution $\theta_k^{(t+1)} = (\mu_k^{(t+1)}, \Sigma_k^{(t+1)})^T$ of equation 5.1.14 or equation 5.1.15 is specified by equation 5.1.17 and equation 5.1.18 when each $f_k(Y|\theta_k)$ is multivariate normal distribution within normal mixture model. The two steps of GEM algorithm for constructing normal mixture model at some t^{th} iteration are refined in table 5.1.3 (Bilmes, 1998, p. 7).

E-step:

The conditional probability $P(k \mid Y_i, \Theta^{(t)})$ is calculated based on current parameter $\Theta^{(t)} = (\alpha_1^{(t)}, \alpha_2^{(t)},\ldots, \alpha_K^{(t)}, \theta_1^{(t)}, \theta_2^{(t)},\ldots, \theta_K^{(t)})^T$, according to equation 5.1.10. Note, in normal mixture model, each observational PDF $f_k(Y|\theta_k)$ is (multivariate) normal distribution with mean vector μ_k and covariance matrix Σ_k such that $\theta_k = (\mu_k, \Sigma_k)^T$.

$$P\left(k\middle|Y_i,\Theta^{(t)}\right)=\frac{\alpha_k^{(t)}f_k\left(Y_i\middle|\theta_k^{(t)}\right)}{\sum_{l=1}^{K}\alpha_l^{(t)}f_l\left(Y_i\middle|\theta_l^{(t)}\right)}$$

M-step:

The next parameter $\Theta^{(t+1)} = (\alpha_1^{(t+1)}, \alpha_2^{(t+1)},\ldots, \alpha_K^{(t+1)}, \theta_1^{(t+1)}, \theta_2^{(t+1)},\ldots, \theta_K^{(t+1)})^T$, which is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with subject to Θ, is calculated by equation 5.1.13, equation 5.1.17, and equation 5.1.18 with current parameter $\Theta^{(t)}$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

$$\alpha_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} P\left(k\middle|Y_i,\Theta^{(t)}\right)$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} P\left(k\middle|Y_i,\Theta^{(t)}\right)Y_i}{\sum_{i=1}^{N} P\left(k\middle|Y_i,\Theta^{(t)}\right)}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{N} P\left(k\middle|Y_i,\Theta^{(t)}\right)\left(\left(Y_i-\mu_k^{(t+1)}\right)\left(Y_i-\mu_k^{(t+1)}\right)^T\right)}{\sum_{i=1}^{N} P\left(k\middle|Y_i,\Theta^{(t)}\right)}$$

Table 5.1.3. E-step and M-step of GEM algorithm for constructing normal mixture model

GEM algorithm converges at some t^{th} iteration. At that time, $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ is the optimal estimate of normal mixture model.

6. Discussions

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 2.8 becomes

$$Q(\Theta'|\Theta) = E\left(\log\left(f(X|\Theta')\right)\middle|Y,\Theta\right) = \sum_{\varphi^{-1}(Y)} k(X|Y,\Theta)\log\left(f(X|\Theta')\right)$$

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^{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 Θ^*. Table 6.2 is the proposal of quasi-PSO-GEM algorithm.

E-step:

As usual, $Q(\Theta \mid \Theta^{(t)})$ is determined based on current $\Theta^{(t)}$ according to equation 2.8. Actually, $Q(\Theta \mid \Theta^{(t)})$ is formulated as function of Θ.

M-step includes four sub-steps:

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 $\Theta^{(t)}$:
$$V_i^{(t+1)} = \omega V_i^{(t)} + r\phi_1\left(\Phi_i^{(t)} - \Psi_i^{(t)}\right) + r\phi_2\left(\Theta^{(t)} - \Psi_i^{(t)}\right) \quad (6.1)$$
Where ω, ϕ_1, and ϕ_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 $V_i^{(t)}$:
$$\Psi_i^{(t+1)} = \Psi_i^{(t)} + V_i^{(t)} \quad (6.2)$$
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 $\Phi_i^{(t+1)} = \Phi_i^{(t)}$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 October 2020 doi:10.20944/preprints201802.0131.v6

4. The next parameter $\Theta^{(t+1)}$ is the swarm's new best position over the best positions of all particles:

$$\Theta^{(t+1)} = \underset{\Phi_i^{(t)}}{\operatorname{argmax}} \left\{ Q\left(\Phi_1^{(t)} \middle| \Theta^{(t)}\right), Q\left(\Phi_2^{(t)} \middle| \Theta^{(t)}\right), \ldots, Q\left(\Phi_n^{(t)} \middle| \Theta^{(t)}\right) \right\} \quad (6.3)$$

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 6.1. 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 6.3 is an approximation of Θ^* 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 of 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

Bilmes, J. A. (1998). *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.* International Computer Science Institute, Department of Electrical Engineering and Computer Science. Berkeley: University of Washington. Retrieved from http://melodi.ee.washington.edu/people/bilmes/mypubs/bilmes1997-em.pdf

Borman, S. (2004). *The Expectation Maximization Algorithm - A short tutorial.* University of Notre Dame, Department of Electrical Engineering. South Bend, Indiana: Sean Borman's Home Page.

Burden, R. L., & Faires, D. J. (2011). *Numerical Analysis* (9th Edition ed.). (M. Julet, Ed.) Brooks/Cole Cengage Learning.

Collins, M., & Barzilay, R. (2005). *Advanced Natural Language Processing - The EM Algorithm.* Massachusetts Institute of Technology, Electrical Engineering and

Computer Science. MIT OpenCourseWare. Retrieved October 9, 2020, from https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-864-advanced-natural-language-processing-fall-2005/lecture-notes/lec5.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

Montgomery, D. C., & Runger, G. C. (2003). *Applied Statistics and Probability for Engineers* (3rd Edition ed.). New York, NY, USA: John Wiley & Sons, Inc.

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

Rosen, K. H. (2012). *Discrete Mathematics and Its Applications* (7nd Edition ed.). (M. Lange, Ed.) McGraw-Hill Companies.

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. (2014, August 4). *Karush–Kuhn–Tucker conditions*. (Wikimedia Foundation) Retrieved November 16, 2014, from Wikipedia website: http://en.wikipedia.org/wiki/Karush–Kuhn–Tucker_conditions

Wikipedia. (2014, October 10). *Set (mathematics)*. (A. Rubin, Editor, & Wikimedia Foundation) Retrieved October 11, 2014, from Wikipedia website: http://en.wikipedia.org/wiki/Set_(mathematics)

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

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.

 doi:10.20944/preprints201802.0131.v6

Contents