**Expectation Maximization Algorithm with Combinatorial Assumption**

Loc Nguyen

Independent scholar, Vietnam

Email: ng\_phloc@yahoo.com

Homepage: www.locnguyen.net

**Abstract**

Expectation maximization (EM) algorithm is a popular and powerful mathematical method for parameter estimation in case that there exist both observed data and hidden data. The EM process depends on an implicit relationship between observed data and hidden data which is specified by a mapping function in traditional EM and a joint probability density function (PDF) in practical EM. However, the mapping function is vague and impractical whereas the joint PDF is not easy to be defined because of heterogeneity between observed data and hidden data. The research aims to improve competency of EM by making it more feasible and easier to be specified, which removes the vagueness. Therefore, the research proposes an assumption that observed data is the combination of hidden data which is realized as an analytic function where data points are numerical. In other words, observed points are supposedly calculated from hidden points via regression model. Mathematical computations and proofs indicate feasibility and clearness of the proposed method which can be considered as an extension of EM.

**Keywords:** expectation maximization (EM), observed data, hidden data, mapping function, joint probability density function, combinatorial assumption, regression model.

**1. Introduction**

Expectation maximization (EM) algorithm developed by Dempster, Laird, and Rubin called DLR (Dempster, Laird, & Rubin, 1977) is an extension of maximum likelihood estimation (MLE) method when there exist both observed data *X* and hidden data *Y* and there is an implicit mapping *φ*: ***X*** → ***Y*** such that *φ*–1(*Y*) = {: *φ*(*X*) = *Y*}. Let *f*(*X* | Θ) be the density probability function (PDF) of random variable *X* and let *g*(*Y* | Θ) be the PDF of random variable *Y*. Note, Θ is (vector) parameter.

The conditional PDF of *X* given *Y*, denoted *k*(*X* | *Y*, Θ), is specified as follows:

Given sample = {*Y*1, *Y*2,…, *YN*} whose all *Yi* (s) are mutually independent and identically distributed (iid), EM has many iterations and each iteration has two steps in which expectation step called E-step determines the conditional expectation *Q*(Θ | Θ(*t*)) and maximization step (M-step) re-estimates parameter as follows:

*E-step*:

The expectation *Q*(Θ | Θ(*t*)) is determined based on current parameter Θ(*t*) (Nguyen, 2020).

|  |  |
| --- | --- |
|  | (1.1) |

*M-step*:

The next parameter Θ(*t*+1) is a maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ. Note that Θ(*t*+1) will become current parameter at the next iteration (the (*t*+1)th iteration).

EM converges at some *t*th iteration; at that time, Θ*\** = Θ(*t*+1) = Θ(*t*) is the optimal estimate of EM process. The expectation *Q*(Θ | Θ(*t*)) based on the mapping *φ*(*X*) represents the traditional EM proposed by DLR. Because it is too vague to specify the mapping *φ*(*X*), practical EM issues the joint PDF of *X* and *Y* denoted *f*(*X*, *Y* | Θ) as prerequisite condition to run EM such that:

The expectation *Q*(Θ | Θ(*t*)) becomes (Nguyen, 2020):

|  |  |
| --- | --- |
|  | (1.2) |

However, if *X* and *Y* are too different in context due to data heterogeneity and they are not independent, it will be difficult to define the PDF *f*(*X*, *Y* | Θ). In general, it is not easy to specify both the mapping *φ*(*X*) in traditional EM and the joint PDF *f*(*X*, *Y* | Θ) in practical EM. Therefore, the purpose of this research is to extend competency of EM, in which the observed data *Y* is assumed to be the combination of *X*, called combinatorial assumption (CA). In other words, *Y* can be calculated by an analytic function of *X*, which is more feasible than specifying the mapping *φ*(*X*) and easier than specifying the joint PDF *f*(*X*, *Y*). The analytic function is arbitrary but it is linear called regression function in this research for convenience and feasibility. In some cases, it is possible to convert some analytic functions into linear functions. The next section is the main one which focuses on EM with CA. In general, this research is an extension of EM algorithm.

**2. EM with combinatorial assumption**

Given sample = {*y*1, *y*2,…, *yN*} of size *N* in which all *Yi* (s) are mutually independent and identically distributed (iid). Suppose *X* = (*x*1, *x*2,…, *xn*) which is vector of size *n* distributes normally with mean vector *μ* and covariance matrix Σ as follows:

|  |  |
| --- | --- |
|  | (2.1) |

Where the superscript “*T*” denotes vector (matrix) transposition operator. The *n*x1 mean *μ* and the *n*x*n* covariance matrix Σ:

Note, *σij*2 is covariance of *xi* and *xj* whereas *σi*2 is variance of *xi*. Let *Y* = (*y*1, *y*2,…, *ym*)*T* be random variable representing all sample random variable *Yi* = (*yi*1, *yi*2,…, *yim*)*T*. Note, *X* is *n*x1 vector and *Y* is *m*x1 vector. Suppose there is an assumption that *Y* is a combination of partial random variables (components) of *X* such that:

As a generalization, let *A* be *m*x*n* matrix whose elements are called regressive coefficients as follows:

This implies

|  |  |
| --- | --- |
|  | (2.2) |

As a convention, let:

Then

The equation above is regression function in which *Y* is called responsor and *X* is called regressor whereas *A* is called regressive matrix. The assumption is combinatorial assumption (CA) aforementioned and the method proposed here is called CA method or CA algorithm. Suppose *Y* distributes normally with mean and covariance matrix *S* as follows:

|  |  |
| --- | --- |
|  | (2.3) |

The *m*x*m* covariance matrix *S* is:

Note, *sij*2 is covariance of *yi* and *yj* whereas *si*2 is variance of *yi*. As a convention, let:

Then

As a convention, is the one which is like *Si* except that *S* is replaced by *S*–1. Note, is vector. Similarly, notations and imply such a meaning. The marginal PDF of *Y* is now defined by support of regression model as follows:

Where parameter Θ = (*μ*, Σ, *A*, *S*)*T* is compound parameter. The equation above is not real total probability rule but it implies that the conditional PDF *f*(*Y* | *X*) is substituted by regression model. Consequently, the expectation *Q*(Θ | Θ(*t*)) becomes:

|  |  |
| --- | --- |
|  | (2.4) |

It is necessary to specify the conditional PDF *f*(*X* | *Y*, Θ). Indeed, we have:

The joint PDF *f*(*X*, *Y* | Θ) which is the numerator of *f*(*X* | *Y*, Θ) is defined:

Where,

The expression is approximated with *μ* as follows:

As a result, *f*0(*X*, *Y* | Θ) and *f*(*X*, *Y* | Θ) is approximated as follows:

And

The approximation by removing *X*-dependency from the expression is reasonable because the PDF *f*(*X* | *Y*, Θ) contains second-order proportion with the built-in expression and this PDF also reflects regression model with another built-in expression including parameter *S*2. In other words, the dependency of on *X* is unnecessary. Moreover, EM process will adjust parameters by the best way later. In following proofs and computations, we will see that such dependency removal also makes the research easy to apply shifted Gaussian integral.

The denominator of *f*(*X* | *Y*, Θ) which is *f*(*Y* | Θ) is the integral of *f*(*X*, *Y* | Θ) over *X*:

Where *B* is the integral of *f*0(*X*, *Y* | Θ) over *X*:

It requires to calculate *B* to determine *f*(*X* | *Y*, Θ). By referring the appendix, we can denote:

Obviously, *B* is totally determined. Thus, *f*(*Y* | Θ) is approximated as follows:

As a result, the PDF *f*(*X* | *Y*, Θ) is approximated as follows:

Let *k*(*Y*|Θ) be the constant with subject to *X* but it is function of *Y* with parameter Θ, which is defined as follows:

|  |  |
| --- | --- |
|  | (2.5) |

Shortly, the conditional PDF *f*(*X* | *Y*, Θ(*t*)) is specified (approximated) at E-step of some *t*th iteration process as follows:

|  |  |
| --- | --- |
|  | (2.6) |

Consequently, the expectation *Q*(Θ | Θ(*t*)) at E-step of some *t*th iteration is totally determined. At M-step of current *t*th iteration, *Q*(Θ|Θ(*t*)) is maximized by setting its partial derivatives regarding Θ to be zero. The first-order partial derivative of *Q*(Θ | Θ(*t*)) with regard to *μ* with note that *Q*(Θ | Θ(*t*)) is analytic function is:

By referring to the appendix, we have:

Note, Σ is invertible and symmetric. The next parameter *μ*(*t*+1) at M-step of some *t*th iteration that maximizes *Q*(Θ|Θ(*t*)) is solution of the equation as follows:

|  |  |
| --- | --- |
|  | (2.7) |

The first-order partial derivative of *Q*(Θ | Θ(*t*)) with regard to Σ with note that *Q*(Θ | Θ(*t*)) is analytic function is:

The next parameter Σ(*t*+1) at M-step of some *t*th iteration that maximizes *Q*(Θ|Θ(*t*)) is solution of equation formed by setting to zero matrix. Let (**0**) denote zero matrix.

We have:

Note, *μ* is replaced by *μ*(*t*). Thus, the next parameter Σ(*t*+1) at M-step of some *t*th iteration that maximizes *Q*(Θ|Θ(*t*)) is obtained:

By referring to the appendix, we obtain:

|  |  |
| --- | --- |
|  | (2.8) |

The first-order partial derivative of *Q*(Θ | Θ(*t*)) with regard to *A*0 with note that *Q*(Θ | Θ(*t*)) is analytic function is:

By referring to the appendix, we obtain:

Note, is replaced by at current *t*th iteration. Therefore, the next parameter *A*0(*t*+1) at M-step of some *t*th iteration that maximizes *Q*(Θ|Θ(*t*)) is obtained by setting the partial derivative to be zero:

|  |  |
| --- | --- |
|  | (2.9) |

Note, the next parameter *μ*(*t*+1) is specified by equation 2.7. The first-order partial derivative of *Q*(Θ | Θ(*t*)) with regard to with note that *Q*(Θ | Θ(*t*)) is analytic function is (Saliba, 2016):

Where denotes Kronecker product and *Im* is *m*x*m* identity matrix. The next parameter at M-step of some *t*th iteration that maximizes *Q*(Θ|Θ(*t*)) is obtained by setting the partial derivative to be zero. Because the Kronecker product occurs, the equation being equal to zero is equivalent to *m*–1 equations and *n* equations , where

And

The partial derivative is

Please see the appendix to know that *E*(*xj* | *f*(*X* | *Yi*, Θ(*t*)) only depends on . The quantity is approximated with *xj* = *μj* such that:

Therefore, the next parameters is determined (defined) as solution of only *m* equations with suppose that *μj* (s) are not zero at all.

|  |  |
| --- | --- |
|  | (2.10) |

The equation 2.10 can be solved by Newton-Raphson method. If *m* < *n*, we lose *n*–*m* equations and so, problem of information loss happens. The first-order partial derivative of *Q*(Θ | Θ(*t*)) with regard to *S* with note that *Q*(Θ | Θ(*t*)) is analytic function is:

The next parameter *S*(*t*+1) at M-step of some *t*th iteration that maximizes *Q*(Θ|Θ(*t*)) is solution of equation formed by setting to zero matrix. We have:

Therefore, we obtain:

|  |  |
| --- | --- |
|  | (2.11) |

Note, *μ*, *A*0 and are replaced by *μ*(*t*), *A*0(*t*) and , respectively. In general, CA method is EM process with two steps as follows:

*E-step*:

Determining the conditional PDF *f*(*X* | *Y*, Θ(*t*)) specified by equation 2.6 based on current parameter Θ(*t*).

*M-step*:

Calculating next parameters Θ(*t*+1) = (*μ*(*t*+1), Σ(*t*+1), *A*(*t*+1), *S*(*t*+1))*T* based on *f*(*X* | *Y*, Θ(*t*)) determined in the E-step, specified by equations 2.7, 2.8, 2.9, 2.10, and 2.11.

|  |
| --- |
|  |

In practice, it is not necessary to compute the covariance matrix Σ(*t*+1) and the variance *S*(*t*+1) because computational cost is high and it is also really ineffective to estimate Σ and *S*. Note, the condition that both Σ and *S* are invertible along with |Σ| ≠ 0 and |*S*| ≠ 0 is not easy to assert over many computational iterations. The most important parameters are *μ* and *A* and we should fix the other parameters Σ and *S* with hints of predefined bias or background knowledge.

**3. Discussions and conclusions**

Although combinatorial assumption (CA) is subjective assumption, it will reach high usefulness and high effectiveness if there is strong regressive relationship between observed data and hidden data in many cases. The regression function may not be linear but it is easy to convert nonlinear functions into linear function in some cases. For instance, there are some nonlinear functions as follows:

|  |  |
| --- | --- |
| Logarithm function |  |
| Exponent function |  |
| Product function |  |

Given logarithm function, the transformation is and *uj* = log(*xj*). For exponent function, the transformation is and *v* = log(*y*). For product function, the transformation is , *v* = log(*y*), *uj* = log(*xj*), and *β*0 = log(*α*0).

Recall that the dimensions of *X* and *Y* are *n* and *m*, respectively. Note, if *m* ≥ *n*, there is no information loss which is ideal case of CA method. Otherwise, if *m* < *n*, the important parameters such as *μ* and *A* vary much with large amplitude due to information loss of dimension reduction. The problem is called large-scale variation. Therefore, in practice there should be restrictions on *μ* and *A* specified as constraints *u*(*μ*) ≤ 0 and *v*(*A*) ≤ 0; for example, slope of regressive hyperplane specified by the normal vectors , which varies in a predefined interval, is a constraint. The solution is to apply Lagrange duality method into maximizing *Q*(Θ|Θ(*t*)), in which a Lagrange function *l*(Θ, *κ*, *λ* | Θ(*t*)) is defined as sum of *Q*(Θ|Θ(*t*)) and constraints *u*(*μ*) ≤ 0 and *v*(*A*) ≤ 0 as follows:

Note *Q*(Θ|Θ(*t*)) is maximized indirectly by maximizing the Lagrange function *l*(Θ, *κ*, *λ* | Θ(*t*)), in which *κ* ≥ 0and *λ* ≥ 0 called Lagrange multipliers are concerned. Another simple trick to alleviate the large-scale variation of *μ* and *A* is to initialize appropriate values *μ*(0) and *A*(0) at the first iteration of EM process.

An application of CA is to learn Bayesian parameter. According to Bayesian statistics, the parameter Θ is considered as random variable. Given sample *D* = {*X*1, *X*2,…, *XN*) whose all observations *Xi* (s) are iid, the posterior probability of Θ given *D* denoted *P*(Θ | *D*, *ξ*) is calculated based on *D* and the prior probability of Θ denoted *P*(Θ | *ξ*).

Where *ξ* denotes background knowledge about Θ which can be considered sub-parameter but we do not focus on learning *ξ*. Let *X* be random variable representing all *Xi*. The most popular method to learn Bayesian parameter is to use binomial sample and beta distribution. Here CA method considers the parameter Θ as hidden data and random variable *X* as observed data.

Therefore, the posterior probability *P*(Θ | *D*, *ξ*) is the same to the conditional PDF *f*(*X* | *y*, Θ) aforementioned. After EM process is finished, *f*(*X* | *y*, Θ) or *P*(Θ | *D*, *ξ*) is obviously determined.

In general, CA is an extension of EM, whose strong point is feasibility and simplicity with clearness of mathematic formulas. Especially, in the ideal case that dimension of *Y* (*y*) is larger than or equal to dimension of *X*, CA method will result out best estimates because there is no information loss. However, its drawback is the large-scale variation in determining important parameters *μ* and *A* when dimension of *Y* (*y*) is smaller than dimension of *X*. Therefore, in the future I will research Lagrange duality method to maximize the expectation *Q*(Θ|Θ(*t*)) with constraints on *μ* and *A*.

**Appendix**

This appendix focuses on calculating integrals related to the following Gaussian function:

Note, ***x*** = (*x*1, *x*2,…, *xh*)*T* is *h*x1 vector, *A* is *n*x*h* matrix, *B* is *m*x*h* matrix, and *c* is *h*x1 vector whereas *R* is *n*x*n* matrix and *S* is *m*x*m* matrix. Moreover, *R* and *S* are invertible and positive definite. Firstly, we calculate the integral of *f*(*x*).

Diagonalizing *R* and *S* as follows:

Where *n*x*n* matrix *U* and *m*x*m* matrix *V* are orthonormal and *n*x*n* matrix Γ and *m*x*m* matrix Λ are eigenmatrices:

We have (Mathprof, 2013):

Let,

We have:

Let

This implies

Suppose *X* distributes normally with mean *μ* = (*μ*1, *μ*2,…, *μh*)*T*, the product *xjxl* is approximated by *μjμl* and then, the quantities *yi*2 and *zi*2 are re-written as follows:

Let

We have:

The integral is re-written as follows:

Let *w* be the maximum value among *n*, *m*, and *h* and let

As a convention, let

We have (Wikipedia, 2021):

We calculate the integral *E*(*xi* | *f*(*x*|Θ)) as follows (Wikipedia, 2021):

Consequently, the integral *E*(*xi*2 | *f*(*x*|Θ)) as follows (Wikipedia, 2021):

Moreover, the integral *E*(*xixj* | *f*(*x*|Θ)) as follows:

In general, we have:

By extending these integrals with vector *x*, we determine:

**References**

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

Mathprof. (2013, March 22). *Multidimensional Gaussian integral*. (PlanetMath ) Retrieved November 23, 2021, from PlanetMath: https://planetmath.org/multidimensionalgaussianintegral

Nguyen, L. (2020). *Tutorial on EM algorithm.* MDPI. Preprints. doi:10.20944/preprints201802.0131.v8

Saliba, G. (2016, July 21). *Not understanding derivative of a matrix-matrix product*. (Stack Overflow) Retrieved December 06, 2021, from Mathematics Stack Exchange: https://math.stackexchange.com/questions/1866757/not-understanding-derivative-of-a-matrix-matrix-product

Wikipedia. (2021, December 3). *List of integrals of exponential functions*. (Wikimedia Foundation) Retrieved December 6, 2021, from Wikipedia website: https://en.wikipedia.org/wiki/List\_of\_integrals\_of\_exponential\_functions