**Finite mixture model**

Loc Nguyen

Loc Nguyen’s Academic Network, 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. Therefore, EM is appropriate to applications which aim to exploit laten aspects under heterogeneous data. Mixture model is a popular and successful application of EM. This report focuses on introducing EM and mixture model as the essential chapter in this book, which is available in (Nguyen, Tutorial on EM algorithm, 2020, pp. 78-88). I also proposed a special regression model associated with mixture model in which missing values are acceptable.

**Keywords:** Expectation maximization (EM) algorithm, mixture model.

**1. Introduction to expectation maximization algorithm**

Literature of expectation maximization (EM) algorithm in this report 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. The preprint “Tutorial on EM algorithm” (Nguyen, Tutorial on EM algorithm, 2020) by Loc Nguyen is also referred in this report.

Now we skim through an introduction of EM algorithm. Suppose there are two spaces ***X*** and ***Y***, in which ***X*** is *hidden space* whereas ***Y*** is *observed space*. We do not know ***X*** but there is a mapping from ***X*** to ***Y*** so that we can survey ***X***by observing ***Y***. The mapping is many-one function *φ*: ***X*** → ***Y*** and we denote *φ*–1(*Y*) = {: *φ*(*X*) = *Y*} as all such that *φ*(*X*) = *Y*. We also denote ***X***(*Y*) = *φ*–1(*Y*). Let *f*(*X* | Θ) be the probability density function (PDF) of random variable and let *g*(*Y* | Θ) be the PDF of random variable . Note, *Y* is also called observation. Equation 1.1 specifies *g*(*Y* | Θ) as integral of *f*(*X* | Θ) over *φ*–1(*Y*).

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

Where Θ is probabilistic parameter represented as a column vector, Θ= (*θ*1, *θ*2,…, *θr*)*T* in which each *θi* is a particular parameter. If *X* and *Y* are discrete, equation 1.1 is re-written as follows:

According to viewpoint of Bayesian statistics, Θ is also random variable. As a convention, let Ω be the domain of Θ such that 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 Θ= (*μ*, *σ*2)*T*. Note that, Θ can degrades into a scalar as Θ = *θ*. The conditional PDF of *X* given *Y*, denoted *k*(*X* | *Y*, Θ), is specified by equation 1.2.

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

According to DLR (Dempster, Laird, & Rubin, 1977, p. 1), ***X*** is called *complete data* and the term “incomplete data” implies existence of ***X*** and ***Y*** where ***X*** is not observed directly and ***X*** is only known by the many-one mapping *φ*: ***X*** → ***Y***. In general, we only know ***Y***, *f*(*X* | Θ), and *k*(*X* | *Y*, Θ) and so our purpose is to estimate Θ based on such ***Y***, *f*(*X* | Θ), and *k*(*X* | *Y*, Θ). Like MLE approach, EM algorithm also maximizes the likelihood function to estimate Θ but the likelihood function in EM concerns ***Y*** and there are also some different aspects in EM which will be described later. Pioneers in EM algorithm firstly assumed that *f*(*X* | Θ) belongs to exponential family with note that many popular distributions such as normal, multinomial, and Poisson belong to exponential family. Although DLR (Dempster, Laird, & Rubin, 1977) proposed a generality of EM algorithm in which *f*(*X* | Θ) 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 1.3 (Dempster, Laird, & Rubin, 1977, p. 3):

|  |  |
| --- | --- |
|  | (1.3) |

Where *b*(*X*) is a function of *X*, which is called base measure and *τ*(*X*) is a vector function of *X*, which is sufficient statistic. For example, the sufficient statistic of normal distribution is *τ*(*X*) = (*X*, *XXT*)*T*. Equation 1.3 expresses the canonical form of exponential family. Recall that Ω is the domain of Θ such that . Suppose that Ω is a convex set. If Θ is restricted only to Ω then, *f*(*X* | Θ) specifies a *regular exponential family*. If Θ lies in a curved sub-manifold Ω0 of Ω then, *f*(*X* | Θ) specifies a *curved exponential family*. The *a*(Θ) is *partition function* for variable *X*, which is used for normalization.

As usual, a PDF is known as a popular form but its exponential family form (canonical form of exponential family) specified by equation 1.3 looks unlike popular form although they are the same. Therefore, parameter in popular form is different from parameter in exponential family form.

For example, multinormal distribution with theoretical mean *μ* and covariance matrix Σ of random variable *X* = (*x*1, *x*2,…, *xn*)*T* has PDF in popular form is:

Hence, parameter in popular form is Θ = (*μ*, Σ)*T*. Exponential family form of such PDF is:

Where,

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 often exponential family parameter if PDF belongs to exponential family and there is no additional information.

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* | Θ) of hidden space belongs to exponential family. E-step and M-step at the *t*th iteration are described in table 1.1 (Dempster, Laird, & Rubin, 1977, p. 4), in which the current estimate is Θ(*t*), with note that *f*(*X* | Θ) belongs to regular exponential family.

|  |
| --- |
| *E-step*:  We calculate current value *τ*(*t*) of the sufficient statistic *τ*(*X*) from observed *Y* and current parameter Θ(*t*) according to following equation:  *M-step*:  Basing on *τ*(*t*), we determine the next parameter Θ(*t*+1) as solution of following equation:  Note, Θ(*t*+1) will become current parameter at the next iteration ((*t*+1)th iteration). |

**Table 1.1.** E-step and M-step of EM algorithm given regular exponential PDF *f*(*X*|Θ)

EM algorithm stops if two successive estimates are equal, Θ*\** = Θ(*t*) = Θ(*t*+1), at some *t*th iteration. At that time we conclude that Θ*\** is the optimal estimate of EM process. As a convention, the estimate of parameter Θ resulted from EM process is denoted Θ\* instead of in order to emphasize that Θ\* is solution of optimization problem.

For further research, DLR gave a preeminent generality of EM algorithm (Dempster, Laird, & Rubin, 1977, pp. 6-11) in which *f*(*X* | Θ) specifies arbitrary distribution. In other words, there is no requirement of exponential family. They define the conditional expectation *Q*(Θ’ | Θ) according to equation 1.4 (Dempster, Laird, & Rubin, 1977, p. 6).

|  |  |
| --- | --- |
|  | (1.4) |

If *X* and *Y* are discrete, equation 2.4 can be re-written as follows:

The two steps of generalized EM (*GEM*) algorithm aim to maximize *Q*(Θ | Θ(*t*)) at some *t*th iteration as seen in table 1.2 (Dempster, Laird, & Rubin, 1977, p. 6).

|  |
| --- |
| *E-step*:  The expectation *Q*(Θ | Θ(*t*)) is determined based on current parameter Θ(*t*), according to equation 1.4. Actually, *Q*(Θ | Θ(*t*)) is formulated as function of Θ.  *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). |

**Table 1.2.** E-step and M-step of GEM algorithm

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

It is deduced from E-step and M-step that *Q*(Θ | Θ(*t*)) is increased after every iteration. How to maximize *Q*(Θ|Θ(*t*)) is the optimization problem which is dependent on applications. For example, the estimate Θ(*t*+1) can be solution of the equation created by setting the first-order derivative of *Q*(Θ|Θ(*t*)) regarding Θ to be zero, *DQ*(Θ|Θ(*t*)) = **0***T*. If solving such equation is too complex or impossible, 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).

In practice, if *Y* is observed as particular *N* observations *Y*1, *Y*2,…, *YN*. Let = {*Y*1, *Y*2,…, *YN*} be the observed sample of size *N* with note that all *Yi* (s) are mutually independent and identically distributed (iid). Given an observation *Yi*, there is an associated random variable *Xi*. All *Xi* (s) are iid and they are not existent in fact. Each is a random variable like *X*. Of course, the domain of each *Xi* is ***X***. Let = {*X*1, *X*2,…, *XN*} be the set of associated random variables. Because all *Xi* (s) are iid, the joint PDF of is determined as follows:

Because all *Xi* (s) are iid and each *Yi* is associated with *Xi*, the conditional joint PDF of given is determined as follows:

The conditional expectation *Q*(Θ’ | Θ) given samples ***X*** and ***Y*** is re-written according to equation 1.5.

|  |  |
| --- | --- |
|  | (1.5) |

Equation 1.5 is proved in (Nguyen, Tutorial on EM algorithm, 2020, pp. 45-47). In case that *f*(*X* | Θ) and *k*(*X* | *Yi*, Θ) belong to exponential family, equation 1.5 becomes equation 1.6 with an observed sample = {*Y*1, *Y*2,…, *YN*}.

|  |  |
| --- | --- |
|  | (1.6) |

Where,

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

The PDF of *Y* becomes:

The PDF *f*(*Y*|Θ) is equivalent to the PDF *g*(*Y*|Θ) mentioned in equation 1.1. Although there is no explicit mapping from ***X*** to ***Y***, the PDF of *Y* above implies an implicit mapping from ***Z*** to ***Y***. The conditional PDF of *X* given *Z* is specified according to Bayes’ rule as follows:

The conditional PDF *f*(*X*|*Y*, Θ) is equivalent to the conditional PDF *k*(*X*|*Y*, Θ) mentioned in equation 1.2. Of course, given *Y*, we always have:

Equation 1.7 specifies the conditional expectation *Q*(Θ’ | Θ) in case that there is no explicit mapping from ***X*** to ***Y*** but there exists the joint PDF of *X* and *Y* (Nguyen, Tutorial on EM algorithm, 2020, p. 48).

|  |  |
| --- | --- |
|  | (1.7) |

Where,

Note, ***X*** is separated from ***Y*** and the complete data ***Z*** = (***X***, ***Y***) is composed of ***X*** and ***Y***. For equation 1.7, the existence of the joint PDF *f*(*X*, *Y* | Θ) can be replaced by the existence of the conditional PDF *f*(*Y*|*X*, Θ) and the prior PDF *f*(*X*|Θ) due to:

In applied statistics, equation 1.4 is often replaced by equation 1.7 because specifying the joint PDF *f*(*X*, *Y* | Θ) is more practical than specifying the mapping *φ*: ***X*** → ***Y***. However, equation 1.4 is more general equation 1.7 because the requirement of the joint PDF for equation 1.7 is stricter than the requirement of the explicit mapping for equation 1.4. In case that *X* and *Y* are discrete, equation 1.7 becomes:

In practice, suppose *Y* is observed as a sample = {*Y*1, *Y*2,…, *YN*} of size *N* with note that all *Yi* (s) are mutually independent and identically distributed (iid). The observed sample is associated with a a hidden set (latent set) = {*X*1, *X*2,…, *XN*} of size *N*. All *Xi* (s) are iid and they are not existent in fact. Let be the random variable representing every *Xi*. Of course, the domain of *X* is ***X***. Equation 1.8 specifies the conditional expectation *Q*(Θ’ | Θ) given such (Nguyen, Tutorial on EM algorithm, 2020, p. 52).

|  |  |
| --- | --- |
|  | (1.8) |

Equation 1.8 is a variant of equation 1.5 in case that there is no explicit mapping between *Xi* and *Yi* but there exists the same joint PDF between *Xi* and *Yi*. If both *X* and *Y* are discrete, equation 1.8 becomes:

|  |  |
| --- | --- |
|  | (1.9) |

If *X* is discrete and *Y* is continuous such that *f*(*X*, *Y* | Θ) = *P*(*X*|Θ)*f*(*Y* | *X*, Θ) then, according to the total probability rule, we have:

Note, when only *X* is discrete, its PDF *f*(*X*|Θ) becomes the probability *P*(*X*|Θ). Therefore, equation 1.10 is a variant of equation 1.8, as follows:

|  |  |
| --- | --- |
|  | (1.10) |

Where *P*(*X* | *Yi*, Θ) is determined by Bayes’ rule, as follows:

Equation 1.10 is the base for estimating the probabilistic mixture model by EM algorithm, which is described in the next section.

**2. Mixture model**

As usual, let ***X*** be the hidden or latent space and let ***Y*** be the observed space. Especially, the random variable *X* in ***X*** represents latent class or latent component of random variable *Y* in ***Y***. Suppose *X* is discrete and ranges in ***X*** = {1, 2,…, *K*}. The so-called probabilisticfinite *mixture model* is represented by the PDF of *Y*, as seen in equation 2.1.

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

Where,

Note, *Y* can be discrete or continuous. Recall that the ultimate purpose of EM algorithm is to maximize *f*(*Y*|Θ) with subject to Θ. Each *fX*(*Y*|*θX*) is called the *X*th partial PDF of *Y* whose partial parameter is *θX*. Each *fX*(*Y*|*θX*) is also called the *X*th observational PDF of *Y*. It is really the conditional PDF of *Y* given *X*, as seen in equation 2.2.

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

From equation 2.1, the mixture model *f*(*Y*|Θ) is the mean of *K* partial PDFs. The variable *X* implies which partial PDF “generates” *Y* (Bilmes, 1998, p. 5).

Each *αX* is called mixture coefficient. It is really the probability of discrete *X*, as seen in equation 2.3. However, in mixture model, each *αX* is also considered as parameter, which is belongs to the compound parameter Θ.

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

The joint probabilistic distribution of *X* and *Y*, which implies the implicit mapping between ***X*** and ***Y***, is product of the mixture coefficient *αX* and the *X*th PDF of *Y*, as seen in equation 2.4.

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

This implies:

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

Equation 2.6 specifies the conditional probability of *X* given *Y*. Please pay attention to this important probability.

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

Following is the proof of equation 2.6. According to Bayes’ rule, we have:

Applying equation 2.3 and equation 2.4, we have:

In other words, equation 2.6 is established■

Now GEM algorithm is applied into mixture model for estimating the parameter Θ. Derived from equation 1.7 in case of discrete *X*, the conditional expectation *Q*(Θ’|Θ) of mixture model becomes:

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

In practice, suppose *Y* is observed as a sample = {*Y*1, *Y*2,…, *YN*} of size *N* in which all *Yi* (s) are mutually independent and identically distributed (iid). The observed sample is associated with a a hidden set (latent set) = {*X*1, *X*2,…, *XN*} of size *N*. All *Xi* (s) are iid and they are not existent in fact. Let be the random variable representing every *Xi*. Of course, the domain of *X* is ***X***. Derived from equation 1.10, equation 2.8 specifies *Q*(Θ’|Θ) given such .

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

Equation 2.8 is the general case of equation 2.7. At the *t*th iteration of GEM, given current parameter Θ(*t*) = (*α*1(*t*), *α*2(*t*),…, *αK*(*t*), *θ*1(*t*), *θ*2(*t*),…, *θK*(*t*))*T*, the conditional expectation specified by equation 2.8 is written as follows:

Thus, the unknown of *Q*(Θ|Θ(*t*)) is Θ = (*α*1, *α*2,…, *αK*, *θ*1, *θ*2,…, *θK*)*T*. Because *X* is discrete and ranges in {1, 2,…, *K*}, the conditional expectation *Q*(Θ|Θ(*t*)) is re-written as equation 2.9 for convenience.

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

Where the conditional probability *P*(*k* | *Y*, Θ(*t*)) is determined by equation 2.10 which is indeed equation 2.6.

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

At M-step of the current *t*th iteration, *Q*(Θ|Θ(*t*)) specified by equation 2.9 is maximized with subject to Θ. How to maximize *Q*(Θ|Θ(*t*)) with subject to Θ is dependent on types of partial PDFs *fk*(*Yi*|*θk*).

Because there is the constraint , we use Lagrange duality method to maximize to maximize *Q*(Θ|Θ(*t*)). The Lagrange function *la*(Θ, *λ* | Θ(*t*)) is sum of *Q*(Θ|Θ(*t*)) and the constraint , which is specified by equation 2.11.

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

Note, *λ* ≥ 0 is called Lagrange multiplier. Of course, *la*(Θ, *λ* | Θ(*t*)) is function of Θ and *λ*. The next parameters *αk*(*t*+1) that maximizes *Q*(Θ|Θ(*t*)) is solution of the equation formed by setting the first-order partial derivative of Lagrange function regarding *αk* and *λ* to be zero with suppose that the Lagrange function is first-order smooth function.

This implies:

|  |  |
| --- | --- |
|  | (2.12) |

Summing equation 2.12 over *K* classes {1, 2,…, *K*}, we have (Bilmes, 1998, p. 5):

Substituting *λ = N* into equation 2.12, the next parameters *αk*(*t*+1) is totally determined by equation 2.13.

|  |  |
| --- | --- |
|  | (2.13) |

Note, the conditional probability *P*(*k* | *Yi*, Θ(*t*)) is determined by equation 2.10.

When parameters *αk*(*t*+1) and *λ* are determined, the Lagrange function *la*(Θ, *λ* | Θ(*t*)) is now function of parameters *θk* as *la*(*θk*|*θk*(*t*)). The next parameters *θk*(*t*+1) is solution of the equation formed by setting the first-order partial derivative of Lagrange function regarding *θk* to be zero with suppose that the Lagrange function is first-order smooth function.

Thus, the next parameters *θk*(*t*+1) is solution of the equation 2.14.

|  |  |
| --- | --- |
|  | (2.14) |

The two steps of GEM algorithm for constructing mixture model at some *t*th iteration are shown in table 2.1. Note, suppose the Lagrange function is first-order smooth function.

|  |
| --- |
| *E-step*:  The conditional probability *P*(*k* | *Yi*, Θ(*t*)) is calculated based on current parameter Θ(*t*) = (*α*1(*t*), *α*2(*t*),…, *αK*(*t*), *θ*1(*t*), *θ*2(*t*),…, *θK*(*t*))*T*, according to equation 2.10.  *M-step*:  The next parameter Θ(*t*+1) = (*α*1(*t*+1), *α*2(*t*+1),…, *αK*(*t*+1), *θ*1(*t*+1), *θ*2(*t*+1),…, *θK*(*t*+1))*T*, which is a maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ, is calculated by equation 2.13 and equation 2.14. Note, *θk*(*t*+1) is solution of the equation 2.14. |

**Table 2.1.** 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, Θ*\** = Θ(*t*+1) = Θ(*t*) is the optimal estimate of mixture model regarding first-order smooth Lagrange function.

Suppose that each PDF *fk*(*Yi*|*θk*) ) belongs to regular exponential family and then, solving equation 2.4 is easier as follows:

(Due to *fk*(*Yi*|*θk*) ) belongs to exponential family)

(Due to log’(*a*(*θk*)) = (*E*(*τ*(*Y*|*θk*)))*T*, please see table 1.2)

In general, the next parameters *θk*(*t*+1) is solution of the equation 2.15 within regular exponential family.

|  |  |
| --- | --- |
|  | (2.15) |

Where *Y* is the random variable representing all *Yi* (s) and,

The two steps of GEM algorithm for constructing mixture model at some *t*th iteration are shown in table 2.2 with suppose that each partial PDF *fX*(*Y*|*θX*) is assumed to belong regular exponential family.

|  |
| --- |
| *E-step*:  The conditional probability *P*(*k* | *Yi*, Θ(*t*)) is calculated based on current parameter Θ(*t*) = (*α*1(*t*), *α*2(*t*),…, *αK*(*t*), *θ*1(*t*), *θ*2(*t*),…, *θK*(*t*))*T*, according to equation 2.10.  *M-step*:  The next parameter Θ(*t*+1) = (*α*1(*t*+1), *α*2(*t*+1),…, *αK*(*t*+1), *θ*1(*t*+1), *θ*2(*t*+1),…, *θK*(*t*+1))*T*, which is a maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ, is calculated by equation 2.13 and equation 2.15. Note, *θk*(*t*+1) is solution of the equation 2.15. |

**Table 2.2.** 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, Θ*\** = Θ(*t*+1) = Θ(*t*) is the optimal estimate of mixture model regarding regular exponential family.

There is a special case that each *fk*(*Yi*|*θk*) 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 *Yi*. Without loss of generality, suppose *Y* is vector so that each *fk*(*Y*|*θk*) is multinormal distribution. Recall that each *fk*(*Y*|*θk*) 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 2.14 or equation 2.15 for *θk*. Suppose random variable *Y* is vector of size *n*.

|  |  |
| --- | --- |
|  | (2.16) |

Where *μk* and Σ*k* are mean vector and covariance matrix of *fk*(*Y*|*θk*), 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 *θk*(*t*+1) = (*μk*(*t*+1), Σ*k*(*t*+1))*T* by solving directly equation 2.14 or equation 2.15. Recall that each *Yi* conforms to multinormal distribution, according to equation 2.16.

Where *μk* and Σ*k* are mean and covariance matrix of *fk*(*Yi*|*θk*), respectively. The Lagrange function is re-written as follows:

Where *p* is the dimension of *Yi*; in other words, *p* is the dimension of space ***Y***.

The first-order partial derivative of Lagrange function with respect to *μk* is (Nguyen, Matrix Analysis and Calculus, 2015, p. 35):

The next parameter *μk*(*t*+1) that maximizes *Q*(Θ|Θ(*t*)) is solution of the equation formed by setting the first-order partial derivative of Lagrange function with regard to *μk* to be **0***T*. Note that **0** = (0, 0,…, 0)*T* is zero vector.

This implies equation 2.17 to specify the next parameter *μk*(*t*+1).

|  |  |
| --- | --- |
|  | (2.17) |

Note, the conditional probability *P*(*k* | *Yi*, Θ(*t*)) is determined by equation 2.10.

The first-order partial derivative of Lagrange function with respect to Σ*k* is:

Due to:

And

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

Where tr(*A*) is trace operator which takes sum of diagonal elements of matrix .

This implies (Nguyen, Matrix Analysis and Calculus, 2015, p. 45):

Where Σ*k* is symmetric and invertible matrix. Substituting the next parameter *μk*(*t*+1) specified by equation 2.16 into the first-order partial derivative of Lagrange function with respect to Σ*k*, we have:

The next parameter Σ*k*(*t*+1) that maximizes *Q*(Θ|Θ(*t*)) is the solution of equation formed by setting the first-order partial derivative of Lagrange function regarding Σ*k* to zero matrix. Let (**0**) denote zero matrix.

We have:

This implies equation 2.18 to specify the next parameter Σ*k*(*t*+1).

|  |  |
| --- | --- |
|  | (2.18) |

Note, the conditional probability *P*(*k* | *Yi*, Θ(*t*)) is determined by equation 2.10 and the next parameter *μk*(*t*+1) is specified by equation 2.17.

As a result, the solution *θk*(*t*+1) = (*μk*(*t*+1), Σ*k*(*t*+1))*T* of equation 2.14 or equation 2.15 is specified by equation 2.17 and equation 2.18 when each *fk*(*Y*|*θk*) is multinormal 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 2.3 (Bilmes, 1998, p. 7).

|  |
| --- |
| *E-step*:  The conditional probability *P*(*k* | *Yi*, Θ(*t*)) is calculated based on current parameter Θ(*t*) = (*α*1(*t*), *α*2(*t*),…, *αK*(*t*), *θ*1(*t*), *θ*2(*t*),…, *θK*(*t*))*T*, according to equation 2.10. Note, in normal mixture model, each observational PDF *fk*(*Y*|*θk*) is (multivariate) normal distribution with mean vector *μk* and covariance matrix Σ*k* such that *θk* = (*μk*, Σ*k*)*T*.  *M-step*:  The next parameter Θ(*t*+1) = (*α*1(*t*+1), *α*2(*t*+1),…, *αK*(*t*+1), *θ*1(*t*+1), *θ*2(*t*+1),…, *θK*(*t*+1))*T*, which is a maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ, is calculated by equation 2.13, equation 2.17, and equation 2.18 with current parameter Θ(*t*). |

**Table 2.3.** E-step and M-step of GEM algorithm for constructing normal mixture model

GEM algorithm converges at some *t*th iteration. At that time, Θ*\** = Θ(*t*+1) = Θ(*t*) is the optimal estimate of normal mixture model.

An interesting application of finite mixture model is soft clustering. Traditional clustering methods assign a fixed cluster to every data point in sample, which means that every data point belongs exactly to one cluster. There are some popular (hard) clustering methods such as *K*-means and *K*-medoids (Han & Kamber, 2006, pp. 451-457). Soft clustering is more flexible when every data point belongs to more than one cluster and the degree of assignment is represented by a probability. Concretely, GEM algorithm for normal mixture model described in table 2.3 is applied into soft clustering. Given sample = {*Y*1, *Y*2,…, *YN*} of size *N* in which all *Yi* (s) are iid and each *Yi* is also called a data point, soft clustering partitions into *K* clusters and each cluster *k* is considered as hidden variable (*X* = 1, 2,…, *K*) and is represented by the aforementioned normal PDF *fk*(*Y*|*θk*)

Where *θk* = (*μk*, Σ*k*)*T* includes mean vector *μk* and covariance matrix Σ*k* of *fk*(*Y*|*θk*), respectively. Especially, *μk* is considered as centroid of cluster *k*. Given cluster *k*, the degree of assignment that a data point *Y* belonging to cluster *k* is specified by such *fk*(*Y*|*θk*). Therefore, GEM algorithm for normal mixture model is used to learn Θ = (*α*1, *α*2,…, *αK*, *θ*1, *θ*2,…, *θK*)*T*. The parameter *αk* indicates degree of popularity of cluster *k*, which can be considered as capacity or size of cluster *k*. It can be also considered as coverage ratio of cluster *k*. The higher the *αk* is, the larger the cluster *k* is. Essentially, soft clustering is to estimate *αk* and *θk* by GEM. Suppose after GEM results out the best estimate Θ\* = (*α*1\*, *α*2\*,…, *αK*\*, *θ*1\*, *θ*2\*,…, *θK*\*)*T*, it is required to determine to which cluster a new data point *Y* is more likely to belong. We calculate *K* joint probabilities *p*1 = *α*1\**f*1(*Y*|*θ*1\*), *p*2 = *α*2\**f*2(*Y*|*θ*2\*),…, and *pK* = *αK*\**fK*(*Y*|*θK*\*). Indeed, each *pk* is the joint probability of *Y* and cluster *k* that come together. Suppose some *pj* is maximum then, *Y* is more likely to belong cluster *j*.

Of course, the probability of each data point *Y* within soft clustering for *K* clusters is

But this probability *f*(*Y*|Θ) is not important. The most important task of GEM for soft clustering is to compute the estimate Θ\* = (*α*1\*, *α*2\*,…, *αK*\*, *θ*1\*, *θ*2\*,…, *θK*\*)*T* from sample in order to determine clusters because each cluster *k* is represented by a pair {*αk*\*, *θk*\*}.

**Example 2.1.** Given sample = {*Y*1, *Y*2, *Y*3, *Y*4}, we apply GEM for soft clustering into *K*=2 clusters.

|  |  |  |
| --- | --- | --- |
|  | *y*1 | *y*2 |
| *Y*1 | 0 | 0 |
| *Y*2 | 0 | 1 |
| *Y*3 | 2 | 0 |
| *Y*4 | 2 | 1 |

Of course, we have *Y*1 = (*y*11=0, *y*12=0)*T*, *Y*2 = (*y*21=0, *y*22=1)*T*, *Y*3 = (*y*31=2, *y*32=0)*T*, and *Y*4 = (*y*41=2, *y*42=1)*T*. The parameter Θ = (*α*1, *α*2, *θ*1, *θ*2)*T* is initialized as follows:

Note, it is easy to calculate normal PDF *fk*(*Y*|*θk*) with known *θk* = (*μk*, Σ*k*)*T*.

At the 1st iteration, E-step we have:

At the 1st iteration, M-step we have:

At the 2nd iteration, E-step we have:

At the 2nd iteration, M-step we have:

Therefore, GEM stops at the 2nd iteration with the estimate Θ(2) = Θ(3) = Θ\* = (*α*1\*, *α*2\*, *θ*1\*, *θ*2\*)*T*.

Given new data point *Y* = (0.5, 0.5)*T*, it is required to determine to which cluster *Y* is more likely to belong. We calculate *K* joint probabilities as follows:

Due to some *p*1=*p*2, the likelihood that *Y* belongs to such two clusters is equal.

**3. Mixture model with dyadic data**

In mixture model, every observation in ordinary sample is univariate or multivariate but there is a case that ordinary sample becomes dyadic sample related to two sets of objects, which causes some modifications of mixture model. *Dyadic data* which is also called co-occurrence data (COD) contains co-occurrent events of objects. It is necessary to obtain statistical models to represent dyadic data and fortunately, finite mixture model is the one. Recall that EM is applied to learn mixture model. Here we focus on EM and mixture model for dyadic data or COD.

Given two finite sets = {*x*1, *x*2,…, *xN*) and = {*y*1, *y*2,…, *yM*) with note that *xi* (s) and *yj* (s) represent -objects and -objects, respectively; exactly, they are names of objects. The numbers of -objects and -objects are =*N* and =*M*, respectively. For example, in information retrieval, *xi* (s) are documents and *yj* (s) are keywords. Hence, *xi* and *yj* are not evaluated as numbers. An observational pair (*xi*, *yj*) is called a *co-occurrence* of *xi* and *yj*. Dyadic data or COD contains these co-occurrences with note that a co-occurrence (*xi*, *yj*) can exist more than one time. So, each co-occurrence (*xi*, *yj*) is indexed by an index *r*. As a result, each co-occurrence is denoted by the triple (*xi*, *yj*, *r*) and we have (Hofmann & Puzicha, 1998, p. 1):

|  |  |
| --- | --- |
|  | (5.1.19) |

Where,

Of course, the size of is . As a convention, *xi*(*r*) and *yj*(*r*) indicate that -object and -object at the *r*th co-occurrence are *xi* and *yj*, respectively. Thus, the triplet (*xi*, *yj*, *r*) can be denoted as (*xi*(*r*), *yj*(*r*), *r*). For example, suppose = {*x*1, *x*2, *x*3) and = {*y*1, *y*2), and dyadic data of 4 co-occurrences, = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*2, 3), (*x*1, *y*1, 4)}, we observe that *x*1 and *y*1 occur together three times at *r*=1, *r*=2, and *r*=4 where as *x*1 and *y*2 occur together one time at *r*=3. In the first co-occurrence (*x*1, *y*1, 1), the notation *x*1(1) indicate that the -object at this co-occurrence is *x*1. In the third co-occurrence (*x*1, *y*2, 3), the notation *y*2(3) indicate that the -object at this co-occurrence is *y*2.

If each co-occurrence of *xi* and *yj* is associated with a value *z* (Hofmann, Puzicha, & Jordan, Learning from Dyadic Data, 1998, p. 1), the triple (*xi*, *yj*, *r*) becomes the quadruplet (*xi*, *yj*, *z*, *r*) which is called *valued co-occurrence* of *xi* and *yj*. The value *z* is called associative value or co-occurrent value. If *z* is value of a variable *Z* then, *Z* is called associative variable or co-occurrent variable. As a result, the sample is called *valued dyadic data* or valued COD. Note, *Z* can be univariate or multivariate (vector).

|  |  |
| --- | --- |
|  | (5.1.20) |

Where,

As a convention, *Z*(*r*) or *z*(*r*) indicates that the associative value at *r*th co-occurrence is *Z*=*z*. Thus, the quadruplet (*xi*, *yj*, *Z*, *r*) can be denoted as (*xi*(*r*), *yj*(*r*), *Z*(*r*), *r*). For example, suppose = {*x*1, *x*2, *x*3) and = {*y*1, *y*2), and valued dyadic sample of 4 co-occurrences, = {(*x*1, *y*1, 6, 1), (*x*1, *y*1, 8, 2), (*x*1, *y*2, 7, 3), (*x*1, *y*1, 9, 4)}, we observe that *x*1 and *y*1 occur together three times at *r*=1, *r*=2, and *r*=4 where as *x*1 and *y*2 occur together one time at *r*=3. Moreover, at *r*=1, *r*=2, *r*=3, and *r*=4, associative values are *Z*(1)=6, *Z*(2)=7, *Z*(3)=8, and *Z*(4)=9, respectively. Valued dyadic data is special case of dyadic data. As a convention, dyadic data is default if there is no additional information.

Given fixed *xk*, let be the -partitioned subset of which contains co-occurrences whose -objects are fixed at *xk* (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 1). Note, can be empty. The size of is .

|  |  |
| --- | --- |
|  | (5.1.21) |

Dyadic data is partitioned into subsets .

Given fixed *yl*, let be the -partitioned subset of which contains co-occurrences whose -objects are fixed at *yl*. Note, can be empty. The size of is .

|  |  |
| --- | --- |
|  | (5.1.22) |

Dyadic data is partitioned into subsets .

Given fixed *xk* and fixed *yl*, let be the subset of the which contains co-occurrences whose -objects and -objects are fixed at *xk* and *yl*. Note, can be empty. The size of is .

|  |  |
| --- | --- |
|  | (5.1.23) |

Let *n*(*xi*) and *n*(*yj*) denote the number of *xi* and the number of *yj*, respectively.

|  |  |
| --- | --- |
|  | (5.1.24) |

Let *n*(*xi*, *yj*) denote the number of *xi* and *yj*.

|  |  |
| --- | --- |
|  | (5.1.25) |

Let *n*(*xi*|*yj*) and *n*(*yj*|*xi*) denote the frequency of *xi* given *yj* and the frequency of *yj* given *xi*, respectively.

|  |  |
| --- | --- |
|  | (5.1.26) |

For example, suppose = {*x*1, *x*2, *x*3) and = {*y*1, *y*2), and dyadic data of 4 co-occurrences, = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*2, 3), (*x*1, *y*1, 4)}, we have = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*2, 3), (*x*1, *y*1, 4)}, = = Ø, = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*1, 4)}, = {(*x*1, *y*2, 3)}, = = {(*x*1, *y*1, 1), (*x*1, *y*1, 2), (*x*1, *y*1, 4)}, = {(*x*1, *y*2, 3)}, = = = = Ø, *n*(*x*1) = 1, *n*(*x*2) = *n*(*x*3) = 0, *n*(*y*1) = 3, *n*(*y*2) = 1, *n*(*x*1, *y*1) = 3, *n*(*x*1, *y*2) = 1, *n*(*x*2, *y*1) = *n*(*x*2, *y*2) = *n*(*x*3, *y*1) = *n*(*x*3, *y*2) = 0, *n*(*x*1 | *y*1) = 1, *n*(*x*1 | *y*2) = 1, *n*(*x*2 | *y*1) = *n*(*x*2 | *y*2) = *n*(*x*3 | *y*1) = *n*(*x*3 | *y*2) = 0, *n*(*y*1 | *x*1) = 3/4, *n*(*y*2 | *x*1) = 1/4.

Suppose each co-occurrence (*xi*, *yj*) belongs to a latent variable *C* and *C* has *K* values *ck* (s). These values *ck* (s) are called classes or aspects and thus, mixture model for dyadic data is also called aspect model or latent class model which aims to discover the latent variable *C*. Without loss of generality, let *ck* = *k* where *k* = 1, 2,…, *K*. The random variable *C* has discrete distribution such that every value has an associated probability *αk*. Of course, there are *K* probabilities *αk* (s). There are three kinds of dyadic mixture model for dyadic data such as symmetric mixture model (SMM), asymmetric mixture model (AMM), and product-space mixture model (PMM). These models were introduced by Hofmann and Puzicha (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998).

The mixture model of dyadic data is called symmetric mixture model (SMM) if *αk* (s) are independent from both *xi* and *yj*. SMM is defined as follows (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 2):

|  |  |
| --- | --- |
|  | (5.1.27) |

Where *αk* is the probability of aspect *k*. Note, *P*(.) denote probability.

The is the probability of *xi* given aspect *k*.

The is the probability of *yj* given aspect *k*.

This implies that *xi* and *yj* are mutually independent in SMM.

The joint probability of *xi*, *yj*, and *k* is:

The parameter of SMM is Θ = (*αk*, *pi*|*k*, *qj*|*k*)*T* in which there are *K*( + + 1) partial parameters *αk*, *pi*|*k*, and *qj*|*k*. Note,

By applying GEM, given dyadic sample , at the *t*th iteration of GEM, given current parameter Θ(*t*) = (*αk*(*t*), *pi*|*k*(*t*), *qj*|*k*(*t*))*T*, the conditional expectation *Q*(Θ|Θ(*t*)) is (Nguyen, Learning Dyadic Data and Predicting Unaccomplished Co-occurrent Values by Mixture Model, 2020, p. 5):

|  |  |
| --- | --- |
|  | (5.1.28) |

Where,

|  |  |
| --- | --- |
|  | (5.1.29) |

Note, *n*(*xi*, *yj*) is the number of co-occurrences (*xi*, *yj*) in , which is specified by equation 5.1.25. Please refer to equation 5.1.6 and equation 5.1.10 to comprehend equation 5.1.29. Because there are three constraints

We use Lagrange duality method to maximize to maximize *Q*(Θ|Θ(*t*)). The Lagrange function *la*(Θ, *λ* | Θ(*t*)) is sum of *Q*(Θ|Θ(*t*)) and these constraints, as follows (Nguyen, Learning Dyadic Data and Predicting Unaccomplished Co-occurrent Values by Mixture Model, 2020, p. 5):

Note, *λ* = (*λ*1, *λ*2, *λ*3)*T* where *λ*1≥0, *λ*2≥0, and *λ*3≥0 are called Lagrange multipliers. Of course, *la*(Θ, *λ* | Θ(*t*)) is function of Θ and *λ*. The next parameters Θ(*t*+1) that maximizes *Q*(Θ|Θ(*t*)) at M-step of some *t*th iteration is solution of the equation formed by setting the first-order partial derivatives of Lagrange function regarding Θ and *λ* to be zero.

The first-order partial derivative of Lagrange function regarding *αk* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over *K* aspects {1, 2,…, *K*}, we have:

This means the next parameters *αk*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.30) |

The first-order partial derivative of Lagrange function regarding *pi*|*k* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over , we have:

This means the next parameters *pi*|*k*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.31) |

Similarly, the next parameters *qj*|*k*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.32) |

The two steps of GEM algorithm for SMM at some *t*th iteration are shown in table 5.1.4.

|  |
| --- |
| *E-step*:  The conditional probability *P*(*k* | *xi*, *yj*, Θ(*t*)) is calculated based on current parameter Θ(*t*) = (*αk*(*t*), *pi*|*k*(*t*), *qj*|*k*(*t*))*T*, according to equation 5.1.29.  *M-step*:  The next parameter Θ(*t*+1) = (*αk*(*t*+1), *pi*|*k*(*t*+1), *qj*|*k*(*t*+1))*T*, which is a maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ, is calculated by equation 5.1.30, equation 5.1.31, and equation 5.1.32. |

**Table 5.1.4.** E-step and M-step of GEM algorithm for SMM

GEM algorithm converges at some *t*th iteration. At that time, Θ*\** = Θ(*t*+1) = Θ(*t*) is the SMM itself. When SMM is applied into soft clustering, dyadic data is clustered according to blocks and each *αk* is coverage ratio of cluster *k* (aspect *k*).

The mixture model of dyadic data is called asymmetric mixture model (AMM) if *αk* (s) are only independent from *xi* or from *yj*. Without loss of generality, given *αk* (s) are only independent from *yj* (of course, it is dependent on *xi*), AMM is defined as follows (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 3):

|  |  |
| --- | --- |
|  | (5.1.33) |

The *αk*|*i* is the probability of aspect *k* given *xi*.

Where *pi* is the probability of *xi*.

The *qj*|*k* is the conditional probability of *yj* given aspect *k*. Suppose *yj* is dependent from *xi* given *k*, we have:

Note, *qj*|*i* is the conditional probability of *yj* given *xi*, which is defined as follows:

The joint probability of *xi*, *yj*, and *k* is:

The parameter of AMM is Θ = (*αk*|*i*, *pi*, *qj*|*k*)*T* in which there are *K*( + ) + partial parameters *αk*|*i*, *pi*, and *qj*|*k*. Note,

By applying GEM, given dyadic sample , at the *t*th iteration of GEM, given current parameter Θ(*t*) = (*αk*(*t*), *pi*|*k*(*t*), *qj*|*k*(*t*))*T*, the conditional expectation *Q*(Θ|Θ(*t*)) is:

|  |  |
| --- | --- |
|  | (5.1.34) |

Where,

|  |  |
| --- | --- |
|  | (5.1.35) |

Please refer to equation 5.1.6 and equation 5.1.10 to comprehend equation 5.1.35. Because there are three constraints

We use Lagrange duality method to maximize to maximize *Q*(Θ|Θ(*t*)). The Lagrange function *la*(Θ, *λ* | Θ(*t*)) is sum of *Q*(Θ|Θ(*t*)) and these constraints, as follows:

Note, *λ* = (*λ*1, *λ*2, *λ*3)*T* where *λ*1≥0, *λ*2≥0, and *λ*3≥0 are called Lagrange multipliers. Of course, *la*(Θ, *λ* | Θ(*t*)) is function of Θ and *λ*. The next parameters Θ(*t*+1) that maximizes *Q*(Θ|Θ(*t*)) at M-step of some *t*th iteration is solution of the equation formed by setting the first-order partial derivatives of Lagrange function regarding Θ and *λ* to be zero.

The first-order partial derivative of Lagrange function regarding *αk* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over *K* aspects {1, 2,…, *K*}, we have:

This means the next parameters *αk*|*i*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.36) |

The first-order partial derivative of Lagrange function regarding *pi* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over , we have:

This means the next parameters *pi*|*k*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.37) |

The first-order partial derivative of Lagrange function regarding *qj*|*k* is:

Setting this partial derivative to be zero, we obtain:

Summing the equation above over , we have:

This means the next parameters *qj*|*k*(*t*+1) is:

|  |  |
| --- | --- |
|  | (5.1.38) |

The two steps of GEM algorithm for AMM at some *t*th iteration are shown in table 5.1.5.

|  |
| --- |
| *E-step*:  The conditional probability *P*(*k* | *xi*, *yj*, Θ(*t*)) is calculated based on current parameter Θ(*t*) = (*αk*|*i*(*t*), *pi*(*t*), *qj*|*k*(*t*))*T*, according to equation 5.1.35.  *M-step*:  The next parameter Θ(*t*+1) = (*αk*|*i*(*t*+1), *pi*(*t*+1), *qj*|*k*(*t*+1))*T*, which is a maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ, is calculated by equation 5.1.36, equation 5.1.37, and equation 5.1.38. |

**Table 5.1.5.** E-step and M-step of GEM algorithm for AMM

GEM algorithm converges at some *t*th iteration. At that time, Θ*\** = Θ(*t*+1) = Θ(*t*) is the AMM itself. When AMM is applied into soft clustering, dyadic data is clustered vertically (horizontally) and each *αk*|*i* is coverage ratio of cluster *k* (aspect *k*) according to *xi*. Soft clustering with AMM is also called one-side clustering.

Product-space mixture model (PMM) is derived from SMM with a minor change that the aspect set {1, 2,…, *K*} is Cartesian product of -aspect set {1, 2,…, } and -aspect set {1, 2,…, }. In other words, the aspect space is still symmetric but is checked (stripped) according to two directions and .

|  |  |
| --- | --- |
|  | (5.1.39) |

For every *k* belongs to {1, 2,…, *K*}, there always exists a respective pair: and . However, for each or each , there are many respective *k* (Nguyen, Learning Dyadic Data and Predicting Unaccomplished Co-occurrent Values by Mixture Model, 2020, p. 10).

|  |  |
| --- | --- |
|  | (5.1.40) |

The sign “” denotes correspondence. For example, given aspect set {1, 2, 3, 4, 5, 6}, -aspect set {*a*, *b*, *c*} and -aspect set {*A*, *B*}, we have a set of six correspondences: 1{*a*, *A*}, 2{*a*, *B*}, 3{*a*, *C*}, 4{*b*, *A*}, 5{*b*, *B*}, 6{*b*, *C*}. Given *a* {*a*, *b*, *c*}, we have three correspondences among *a* and aspect set {1, 2, 3, 4, 5, 6} such as *a*1, *a*2, and *a*3.

PMM is defined as follows (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 4):

|  |  |
| --- | --- |
|  | (5.1.41) |

As usual, *αk* is the probability of aspect *ck* but is the probability of *xi* given of *k* and is the probability of *yj* given of *k*.

The joint probability of *xi*, *yj*, and *k* is:

The parameter of PMM is Θ = (*αk*, , )*T* in which there are *K* + + partial parameters *αk*, , and . Note,

Learning PMM is like learning SMM and so it is not necessary to duplicate the expansion of *Q*(Θ|Θ(*t*)). The two steps of GEM algorithm for PMM at some *t*th iteration are shown in table 5.1.6.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *E-step*:  The conditional probabilities *P*(*k* | *xi*, *yj*, Θ(*t*)), *P*( | *xi*, *yj*, Θ(*t*)), and *P*( | *xi*, *yj*, Θ(*t*)) are calculated based on current parameter Θ(*t*) = , according to equation 5.1.42, equation 5.1.43, and equation 5.1.44 (Nguyen, Learning Dyadic Data and Predicting Unaccomplished Co-occurrent Values by Mixture Model, 2020, p. 10).   |  |  | | --- | --- | |  | (5.1.42) | |  | (5.1.43) | |  | (5.1.44) |   Please refer to equation 5.1.6 and equation 5.1.10 to comprehend equation 5.1.42.  *M-step*:  The next parameter Θ(*t*+1) = , which is the maximizer of *Q*(Θ | Θ(*t*)) with subject to Θ, is calculated by equation 5.1.45, equation 5.1.46, and equation 5.1.47.   |  |  | | --- | --- | |  | (5.1.45) | |  | (5.1.46) | |  | (5.1.47) | |

**Table 5.1.6.** E-step and M-step of GEM algorithm for PMM

GEM algorithm converges at some *t*th iteration. At that time, Θ*\** = Θ(*t*+1) = Θ(*t*) is the PMM itself. When PMM is applied into soft clustering, dyadic data is clustered in checked (stripped) and each *αk* is coverage ratio of cluster *k* (aspect *k*) but such cluster *k* corresponds to a pair of cluster and cluster . Soft clustering with PMM is also called two-side clustering.

When is valued dyadic data in which every co-occurrence (*xi*, *yj*) is associated with value *z* from random variable *Z* then, SMM is reformed as follows (Nguyen, Learning Dyadic Data and Predicting Unaccomplished Co-occurrent Values by Mixture Model, 2020, pp. 11-12):

|  |  |
| --- | --- |
|  | (5.1.48) |

AMM is reformed as follows:

|  |  |
| --- | --- |
|  | (5.1.49) |

PMM is reformed as follows:

|  |  |
| --- | --- |
|  | (5.1.50) |

Where *fk*(*Z*|*φk*) is the *k*th PDF of *Z* corresponding to the aspect *k*, in which *φk* is parameter of *fk*(*Z*|*φk*). Of course, the parameter Θ now must include all *φk*. It is possible to consider that

Moreover, *Z* is only dependent on *k*.

Note, suppose *xi* and *yj* (as well as *yj* given *xi*) are independent from *Z* given aspect *k*, which is the hint to reform these models.

For example, within SMM, the joint PDF of *xi*, *yj*, *Z*, and *k* is:

Within AMM, the joint PDF of *xi*, *yj*, *Z*, and *k* is:

Within PMM, the joint PDF of *xi*, *yj*, *Z*, and *k* is:

Here it is only necessary to estimate *φk* because how to estimate other partial parameters was aforementioned. By reforming the conditional expectation *Q*(Θ|Θ(*t*)), it is easy to find out that the next parameter *φk*(*t*+1) is solution of following equation:

|  |  |
| --- | --- |
|  | (5.1.51) |

Where *P*(*k* | *xi*(*r*), *yj*(*r*), Θ(*t*)) is specified by equation 5.1.29, equation 5.1.35, and equation 5.1.42 for SMM, AMM, and PMM, respectively. Especially, if *fk*(*Z*|*φk*) distributed normally, the next parameter *φk*(*t*+1) = (*μk*(*t*+1), Σ*k*(*t*+1))*T* containing mean *μk*(*t*+1) and covariance matrix Σ*k*(*t*+1) is calculated as follows:

|  |  |
| --- | --- |
|  | (5.1.52) |

Where *P*(*k* | *xi*(*r*), *yj*(*r*), Θ(*t*)) is specified by equation 5.1.29, equation 5.1.35, and equation 5.1.42 for SMM, AMM, and PMM, respectively. Please refer to equation 5.1.17 and equation 5.1.18 to comprehend equation 5.1.52.

**Example 5.1.2.** Suppose = {*x*1, *x*2} and = {*y*1}, and valued dyadic sample of 4 co-occurrences, = {(*x*1, *y*1, 1, 1), (*x*1, *y*1, 9, 2) }, we will learn SMM given by GEM shown in table 5.1.4. Let *Z* be associative variable which distributes normally with mean – variance *φk* = (*μk*, *σk*2)*T* and is learned by equation 5.1.52. Obviously, we have *Z*(1)=1, *Z*(2)=9, *n*(*x*1, *y*1) = 2, and *n*(*x*2, *y*1) = 0. Suppose the number of aspects is *K*=2. The parameter Θ = (*αk*, *pi*|*k*, *qj*|*k*, *φk*)*T* of SMM is initialized as follows:

|  |
| --- |
|  |

At the 1st iteration, E-step, we have:

At the 1st iteration, M-step, we have:

At the 2nd iteration, E-step, we have:

Note, because the probabilities *P*(*k*=1 | *x*2, *y*1, Θ(2)) and *P*(*k*=2 | *x*2, *y*1, Θ(2)) are arbitrary (0/0), they are assigned to be 0.5.

At the 2nd iteration, M-step, we have:

Therefore, GEM stops at the 2nd iteration with the estimate Θ(2) = Θ(3) = Θ\* = (*αk*\*, *pi*|*k*\*, *qj*|*k*\*, *φk*\*)*T*.

|  |
| --- |
|  |

Similarly, it is easy to learn AMM and PMM ■

**4. Mixture Regression Model (MRM)**

This section describes a special regression model associated with mixture model in which missing values are acceptable. Please read my papers “Mixture Regression Model for Incomplete Data” (Nguyen & Shafiq, Mixture Regression Model for Incomplete Data, 2018) and “Fetal Weight Estimation in Case of Missing Data” (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018) to comprehend regression model and incomplete data.

The probabilistic Mixture Regression Model (MRM) is a combination of normal mixture model and linear regression model. In MRM, the probabilistic Entire Regression Model (ERM) is sum of *K* weighted probabilistic Partial Regression Models (PRMs). Equation 12 specifies MRM (Bilmes, 1998, p. 3).

|  |  |
| --- | --- |
|  | (4.1) |

Where,

Note, Θ is called entire parameter,

The superscript “*T*” denotes transposition operator in vector and matrix. In equation 4.1, the probabilistic distribution *P*(*zi*|*Xi*, Θ) represents the ERM where *zi* is the response variable, dependent variable, or outcome variable. The probabilistic distribution *Pk*(*zi*|*Xi*, *αk*, *σk*2) represents the *k*th PRM *zi* = *αk*0 *+ αk*1*xi*1 *+ αk*2*xi*2 *+ … + αknxin* with suppose that each *zi* conforms to normal distribution according to equation 4.2 with mean *μk* = *αkTXi* and variance *σk*2.

|  |  |
| --- | --- |
|  | (4.2) |

The parameter *αk* = (*αk*0, *αk*1,…, *αkn*)*T* is called the *k*th Partial Regression Coefficient (PRC) and *Xi* = (1, *xi*1, *xi*2,…, *xin*)*T* is data vector. Each *xij* in every PRM is called a regressor, predictor, or independent variable.

In equation 4.1, each mixture coefficient *ck* is the prior probability that any *zi* belongs to the *k*th PRM. Let *Y* be random variable representing PRMs, *Y* = 1, 2,…, *K*. The mixture coefficient *ck* is also called the *k*th weight, which is defined by equation 4.3. Of course, there are *K* mixture coefficients, *K* PRMs, and *K* PRCs.

|  |  |
| --- | --- |
|  | (4.3) |

For each *k*th PRM, suppose each has an inverse regression model (IRM) *xij* = *βkj*0 *+ βkj*1*zi*. In other words, *xij* now is considered as the random variable conforming to normal distribution according to equation 4.4 (Lindsten, Schön, Svensson, & Wahlström, 2017, p. 8).

|  |  |
| --- | --- |
|  | (4.4) |

Where *βkj* = (*βkj*0, *βkj*1)*T* is an inverse regression coefficient (IRC) and (1, *zi*)*T* becomes an inverse data vector. The mean and variance of each *xij* with regard to the inverse distribution *Pkj*(*xij*|*zi*, *βkj*) are *βkjT*(1, *zi*)*T* and *τkj*2, respectively. Of course, for each *k*th PRM, there are *n* IRMs *Pkj*(*xij*|*zi*, *βkj*) and *n* associated IRCs *βkj*. Totally, there are *n*\**K* IRMs associated with *n*\**K* IRCs. Suppose IRMs with fixed *j* have the same mixture model as MRM does. Equation 4.5 specifies the mixture model of IRMs.

|  |  |
| --- | --- |
|  | (4.5) |

In this research, we focus on estimating the entire parameter Θ = (*ck*, *αk*, *σk*2, *βkj*)*T* where *k* is from 1 to *K*. In other words, we aim to estimate *ck*, *αk*, *σk*2, and *βkj* for determining the ERM in case of missing data. As a convention, let Θ\* = (*ck*\*, *αk*\*, (*σk*2)\*, *βkj*\*)*T* be the estimate of Θ = (*ck*, *αk*, *σk*2, *βkj*)*T*, respectively. Let ***D*** = (***X***, ***Z***) be collected sample in which ***X*** is a set of regressors and ***Z*** is a set of outcome variables plus values 1, respectively (Lindsten, Schön, Svensson, & Wahlström, 2017, p. 8) with note that both ***X*** and ***Z*** are incomplete. In other words, ***X*** and ***Z*** have missing values. As a convention, let *zi*– and *xij*– denote missing values of ***Z*** and ***X***, respectively.

|  |  |
| --- | --- |
|  | (4.6) |

The expectation of sufficient statistic *zi* regard to the *k*th PRM *Pk*(*zi*|*Xi*, *αk*, *σk*2) is specified by equation 4.7 (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018).

|  |  |
| --- | --- |
|  | (4.7) |

Where *xi*0=1 for all *i*. The expectation of the sufficient statistic *xij* with regard to each IRM *Pkj*(*xij*|*zi*, *βj*) of the *k*th PRM *Pk*(*zi*|*Xi*, *αk*, *σk*2) is specified by equation 4.8 (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018).

|  |  |
| --- | --- |
|  | (4.8) |

Please pay attention to equations 4.7 and 4.8 because missing values of data ***X*** and data ***Z*** will be estimated by these expectations later.

Because ***X*** and ***Z*** are incomplete, we apply expectation maximization (EM) algorithm into estimating Θ\* = (*ck*\*, *αk*\*, (*σk*2)\*, *βkj*\*)*T*. According to (Dempster, Laird, & Rubin, 1977), EM algorithm has many iterations and each iteration has expectation step (E-step) and maximization step (M-step) for estimating parameters. Given current parameter Θ(*t*) = (*ck*(*t*), *αk*(*t*), (*σk*2)(*t*), *βkj*(*t*))*T* at the *t*th iteration, missing values *zi*– and *xij*– are calculated in E-step so that ***X*** and ***Z*** become complete. In M-step, the next parameter Θ(*t*+1) = (*ck*(*t*+1), *αk*(*t*+1), (*σk*2)(*t*+1), *βkj*(*t*+1))*T* is determined based on the complete data ***X*** and ***Z*** fulfilled in E-step. Here we proposed a so-called Mixture Regression Expectation Maximization (MREM) which is the full combination of Regression Expectation Maximization (REM) algorithm (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018) and mixture model in which we use two EM processes in the same loop. Firstly, we use the first EM process for exponential family of probability distributions to estimate missing values in E-step. The technique is the same to the technique of REM in previous research (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018). Secondly, we use the second EM process to estimate Θ\* for full mixture model in M-step.

Firstly, we focus on fulfilling missing values in E-step. The most important problem in our research is how to estimate missing values *zi*– and *xij*–. Recall that, for each *k*th PRM, every missing value *zi*– is estimated as the expectation based on the current parameter *αk*(*t*), according to equation 4.7 (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018).

Note, *xi*0 = 1. Let *Mi* be a set of indices of missing values *xij*– with fixed *i* for each *k*th PRM. In other words, if then, *xij* is missing. The set *Mi* can be empty. The equation 4.7 is re-written for each *k*th PRM as follows (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018):

According to equation 4.8, missing value *xij*– is estimated by (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018):

Combining equation 4.7 and equation 4.8, we have (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018):

It implies (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018):

As a result, equation 4.9 is used to estimate or fulfill missing values for each *k*th PRM (Nguyen & Ho, Fetal Weight Estimation in Case of Missing Data, 2018).

|  |  |
| --- | --- |
|  | (4.9) |

Now in M-step we use EM algorithm again to estimate the next parameter Θ(*t*+1) = (*ck*(*t*+1), *αk*(*t*+1), (*σk*2)(*t*+1), *βkj*(*t*+1))*T* with current known parameter Θ(*t*) = (*ck*(*t*), *αk*(*t*), (*σk*2)(*t*), *βkj*(*t*+1))*T* given data ***X*** and data ***Z*** fulfilled in E-step. The conditional expectation *Q*(Θ|Θ(*t*)) with unknown Θ is determined as follows (Bilmes, 1998, p. 4):

The next parameter Θ(*t*+1) is a constrained optimizer of *Q*(Θ|Θ(*t*)). This is the optimization problem.

By applying Lagrange method, each next mixture coefficient *ck*(*t*+1) is specified by equation 4.10 (Bilmes, 1998, p. 7).

|  |  |
| --- | --- |
|  | (4.10) |

Where *P*(*Y*=*k* | *Xi*, *zi*, *αk*(*t*), (*σk*2)(*t*)) is specified by equation 4.11 (Bilmes, 1998, p. 3). It is the conditional probability of the *k*th PRM given *Xi* and *zi*. Please pay attention to this important probability. The proof of equation 4.11 is found in (Bilmes, 1998, p. 3), according to Bayes’ rule.

|  |  |
| --- | --- |
|  | (4.11) |

Note, *Pk*(*zi*|*Xi*, *αk*(*t*), (*σk*2)(*t*)) is determined by equation 4.2.

By applying Lagrange method, each next regression coefficient *αk*(*t*+1) is solution of equation 4.12 (Bilmes, 1998, p. 7).

|  |  |
| --- | --- |
|  | (4.12) |

Where **0** = (0, 0,…, 0)*T* is zero vector and *P*(*Y*=*k* | *Xi*, *zi*, *αk*(*t*), (*σk*2)(*t*)) is specified by equation 22. Equation 4.12 is equivalent to equation 4.13:

|  |  |
| --- | --- |
|  | (4.13) |

Let,

Note,

The left-hand side of equation 4.13 becomes:

Where ***U***(*t*) is specified by equation 4.14.

|  |  |
| --- | --- |
|  | (4.14) |

Let,

Note,

The right-hand side of equation 4.13 becomes:

Where *Vi*(*t*) is specified by equation 4.15.

|  |  |
| --- | --- |
|  | (4.15) |

Equation 4.3 becomes:

Which is equivalent to the following equation:

As a result, the next regression coefficient *αk*(*t*+1), which is solution of equation 4.12, is specified by equation 4.16.

|  |  |
| --- | --- |
|  | (4.16) |

Where ***X***, ***U***(*t*), and *Vi*(*t*) are specified by equation 4.6, 4.14, and 4.15, respectively. The proposed equation 4.16 is most important in this research because it is the integration of least squares method and mixture model. If we think deeply, it is the key to combine REM and mixture model. In other words, it is the key to combine two EM processes in the same loop.

By applying Lagrange method, each next partial variance (*σk*2)(*t*+1) is specified by equation 4.17 (Bilmes, 1998, p. 7).

|  |  |
| --- | --- |
|  | (4.17) |

Where *P*(*Y*=*k* | *zi*, *αk*(*t*), (*σk*2)(*t*)) is specified by equation 4.11 and *αk*(*t*+1) is specified by equation 4.16. The proof of equations 21, 23, and 28 is found in (Bilmes, 1998, pp. 5-6).

By using maximum likelihood estimation (MLE) method (Lindsten, Schön, Svensson, & Wahlström, 2017, pp. 8-9), we retrieve equation 4.18 to estimate each next IRC *βkj*(*t*+1) (Montgomery & Runger, 2010, p. 457).

|  |  |
| --- | --- |
|  | (4.18) |

Where ***Z*** and *Xj* are specified in equation 17. Not ***Z*** and *Xj* are fulfilled in E-step. In general, MREM is the full combination of REM and mixture model in which two EM processes are applied into the same loop of E-step and M-step. These steps are described in table 1.

|  |
| --- |
| *E-step*: This is the first EM process. Missing values (*zi*–)*k* and (*xij*–)*k* for each *k*th PRM are fulfilled by equation 4.9 given current parameter Θ(*t*). Please pay attention that each *k*th PRM owns a partial complete data (***X****k*, ***Z****k*). In other words, the whole sample (***X***, ***Z***) has *K* versions (***X****k*, ***Z****k*) for *K* PRMs. Note, such *K* versions are changed over each iteration.  The whole sample (***X***, ***Z***) is fulfilled to become complete data when its missing values *zi*–and *xij*– are aggregated from (*zi*–)*k* and (*xij*–)*k* of *K* versions (***X****k*, ***Z****k*).  *M-step*: This is the second EM process. The next parameter Θ(*t*+1) is determined by equations 4.10, 4.16, 4.17, and 4.18 and the complete data (***X***, ***Z***) fulfilled in E-step.  Where ***U***(*t*) and ***V***(*t*) are specified by equations 4.14 and 4.15 and,  The next parameter Θ(*t*+1) becomes current parameter in the next iteration. |

**Table 1.** Mixture Regression Expectation Maximization (MREM) Algorithm.

EM algorithm stops if at some *t*th iteration, we have Θ(*t*) = Θ(*t*+1) = Θ*\**. At that time, Θ\* = (*ck*\*, *αk*\*, (*σk*2)\*, *βkj*\*) is the optimal estimate of EM algorithm. Note, Θ(1) at the first iteration is initialized arbitrarily. Here MREM stops if ratio deviation between Θ(*t*) and Θ(*t*+1) is smaller than a small enough terminated threshold *ε*> 0 or MREM reaches a large enough number of iterations. The smaller the terminated threshold is, the more accurate MREM is. MREM uses both the terminated threshold *ε* = 0.1% = 0.001 and the maximum number of iterations (10000). The maximum number of iterations prevents MREM from running for a long time.

There are some more problems related to missing data inside the mixture regression model that I proposed but this report focuses on the mixture model. Therefore, I suspend discussion on mixture model here. I wish to have another opportunity to present more about mixture model and regression model.

**References**

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

Burden, R. L., & Faires, D. J. (2011). *Numerical Analysis* (9th Edition ed.). (M. Julet, Ed.) Brooks/Cole Cengage Learning.

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.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd Edition ed.). (J. Gray, Ed.) San Francisco, CA, USA: Morgan Kaufmann Publishers, Elsevier.

Lindsten, F., Schön, T. B., Svensson, A., & Wahlström, N. (2017). *Probabilistic modeling – linear regression & Gaussian processes.* Uppsala University, Department of Information Technology. Uppsala: Uppsala University. Retrieved January 24, 2018, from http://www.it.uu.se/edu/course/homepage/sml/literature/probabilistic\_modeling\_compendium.pdf

Montgomery, D. C., & Runger, G. C. (2010). *Applied Statistics and Probability for Engineers* (5th ed.). Hoboken, New Jersey, USA: John Wiley & Sons. Retrieved from https://books.google.com.vn/books?id=\_f4KrEcNAfEC

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

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

Nguyen, L., & Ho, T.-H. T. (2018). Fetal Weight Estimation in Case of Missing Data. (T. Schmutte, Ed.) *Experimental Medicine (EM) - Special Issue "Medicine and Healthy Food"*.

Nguyen, L., & Shafiq, A. (2018, December 31). Mixture Regression Model for Incomplete Data. (L. d. Istael, Ed.) *Revista Sociedade Científica, 1*(3), 1-25. doi:10.5281/zenodo.2528978

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

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