Proceedings of the 6th International Conference on Future of Education, Vol. 6, 2023, pp. 152 - 171

ISSN 2630-7413 online
DOI: https://doi.org/10.17501/26307413.2023.6112

LEARNING DYADIC DATA AND PREDICTING UNACCOMPLISHED CO-OCCURRENT VALUES BY MIXTURE MODEL

Nguyen L[1]* and Lanuza MH[2]

[1]*Loc Nguyen's Academic Network, Vietnam*
[2]*Philippine Normal University, Manila, Philippines*

Abstract: Dyadic data which is also called co-occurrence data (COD) contains co-occurrences of objects where these objects are indexed and grouped into two finite sets. It is necessary to model dyadic data by applied mathematical tools because dyadic data analysis is interesting and important to many applications relating to indexed two-dimensional data such as image processing and recommendation collaborative filtering. Fortunately, finite mixture model is a solid statistical model to learn and make inference on dyadic data because mixture model is built smoothly and reliably by expectation maximization (EM) algorithm which is suitable to inherent spareness of dyadic data. This research summarizes mixture models for dyadic data, in which there are three well-known models such as symmetric mixture model (SMM), asymmetric mixture model (AMM), and product-space mixture model (PMM) which are described by beautiful mathematical proofs and explanations derived from EM algorithm. Objects in traditional dyadic data are indexed as categories and so their potential real values are concerned because of potential applications and extensions of dyadic data analysis. For instance, when each co-occurrence in dyadic data is associated with a real value, there are many unaccomplished values because a lot of co-occurrences are inexistent. In the research, these unaccomplished values are estimated as mean (expectation) of random variable given partial probabilistic distributions inside dyadic mixture model. This estimation result is solid due to support of EM algorithm.

Keywords: dyadic data, co-occurrence data, expectation maximization (EM) algorithm, mixture model.

Introduction

Suppose data has two parts such as hidden part X and observed part Y and we only know Y. A relationship between random variable X and random variable Y is specified by the joint probabilistic density function (PDF) denoted $f(X, Y \mid \Theta)$ where Θ is parameter. Given sample $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$ whose all Y_i (s) are mutually independent and identically distributed (iid), it is required to estimate Θ based on $\mathcal{Y}$ whereas X is unknown. Expectation maximization (EM) algorithm is applied to solve this problem when only $\mathcal{Y}$ is observed. EM has many iterations and each iteration has two steps such as expectation step (E-step) and maximization step (M-step). At some t^{th} iteration, given current parameter $\Theta^{(t)}$, the two steps are described as follows:

*Corresponding Author's Email: ng_phloc@yahoo.com

Table 1.1: E-step and M-step of EM algorithm.

E-step: The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current parameter $\Theta^{(t)}$, according to equation 1.1 (Nguyen, 2020, p. 50). $$Q\left(\Theta\middle|\Theta^{(t)}\right) = \sum_{i=1}^{N} \int_{X} f\left(X\middle|Y_i, \Theta^{(t)}\right) \log\left(f(X, Y_i|\Theta)\right) \mathrm{d}X \quad (1.1)$$ *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 $(t+1)^{th}$ iteration).

EM algorithm will converge after some iterations, at that time we have the estimate $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^{*}$. Note, the estimate Θ^{*} is result of EM. The EM algorithm shown in Table 1.1 is also called general EM or GEM.

Especially, the random variable X represents latent class or latent component of random variable Y. Suppose X is discrete and ranges in $\{1, 2, \ldots, K\}$. As a convention, let $k=X$. Note, because all Y_i (s) are iid, let random variable Y represent every Y_i. The so-called probabilistic finite mixture model is represented by the PDF of Y, as follows:

$$f(Y|\Theta) = \sum_{k=1}^{K} \alpha_k f_k(Y|\theta_k) \quad (1.2)$$

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, the superscript "T" denotes transpose operator for vector and matrix. The $Q(\Theta \mid \Theta^{(t)})$ is re-defined for finite mixture model as follows (Nguyen, 2020, p. 79):

$$Q\left(\Theta\middle|\Theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} P\left(k\middle|Y_i, \Theta^{(t)}\right) \log\left(\alpha_k f_k(Y_i|\theta_k)\right) \quad (1.3)$$

Where,

$$P\left(k\middle|Y_i,\Theta^{(t)}\right)=\frac{\alpha_k^{(t)}f_k\left(Y_i\middle|\theta_k^{(t)}\right)}{\sum_{l=1}^{K}\alpha_l^{(t)}f_l\left(Y_i\middle|\theta_l^{(t)}\right)} \quad (1.4)$$

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. 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. It is easy to recognize that when mixture model is applied into soft clustering, latent class *k* represents a cluster.

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. The next section focuses on mixture model for dyadic data.

Mixture models for dyadic data

Given two finite sets $\mathcal{X}$ = {x_1, x_2,…, x_N) and $\mathcal{Y}$ = {y_1, y_2,…, y_M) with note that x_i (s) and y_j (s) represent $\mathcal{X}$-objects and $\mathcal{Y}$-objects, respectively; exactly, they are names of objects. The numbers of $\mathcal{X}$-objects and $\mathcal{Y}$-objects are $|\mathcal{X}|$=*N* and $|\mathcal{Y}|$=*M*, respectively. For example, in information retrieval, x_i (s) are documents and y_j (s) are keywords. Hence, x_i and y_j are not evaluated as numbers. An observational pair (x_i, y_j) $\in \mathcal{X}\times\mathcal{Y}$ is called a *co-occurrence* of x_i and y_j. Dyadic data or COD $\mathcal{S}$ contains these co-occurrences with note that a co-occurrence (x_i, y_j) can exist more than one time. So, each co-occurrence (x_i, y_j) is indexed by an index *r*. As a result, each co-occurrence is denoted by the triple (x_i, y_j, *r*) and we have (Hofmann & Puzicha, 1998, p. 1):

$$\mathcal{S}=\left\{\left(x_i,y_j,r\right):1\le r\le|\mathcal{S}|\right\} \quad (2.1)$$

Where,

$$x_i\in\mathcal{X}=\left\{x_1,x_2,\dots,x_{|\mathcal{X}|}\right\}$$
$$y_j\in\mathcal{Y}=\left\{y_1,y_2,\dots,y_{|\mathcal{Y}|}\right\}$$

Of course, the size of $\mathcal{S}$ is $|\mathcal{S}|$. As a convention, $x_i(r)$ and $y_j(r)$ indicate that $\mathcal{X}$-object and $\mathcal{Y}$-object at the r^{th} co-occurrence are x_i and y_j, respectively. Thus, the triplet (x_i, y_j, *r*) can be denoted as ($x_i(r)$, $y_j(r)$, *r*). For example, suppose $\mathcal{X}$ = {x_1, x_2, x_3) and $\mathcal{Y}$ = {y_1, y_2), and dyadic data of 4 co-occurrences, $\mathcal{S}$ = {(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 $\mathcal{X}$-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 $\mathcal{Y}$-object at this co-occurrence is y_2.

If each co-occurrence of x_i and y_j is associated with a value z (Hofmann, Puzicha, & Jordan, Learning from Dyadic Data, 1998, p. 1), the triple (x_i, y_j, r) becomes the quadruplet (x_i, y_j, z, r) which is called *valued co-occurrence* of x_i and y_j. 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 $\mathcal{S}$ is called *valued dyadic data* or valued COD. Note, Z can be univariate or multivariate (vector).

$$\mathcal{S} = \{(x_i, y_j, Z, r): 1 \leq r \leq |\mathcal{S}|\} \tag{2.2}$$

Where,

$$x_i \in \mathcal{X} = \{x_1, x_2, \dots, x_{|\mathcal{X}|}\}$$
$$y_j \in \mathcal{Y} = \{y_1, y_2, \dots, y_{|\mathcal{Y}|}\}$$

As a convention, $Z(r)$ or $z(r)$ indicates that the associative value at r^{th} co-occurrence is $Z=z$. Thus, the quadruplet (x_i, y_j, Z, r) can be denoted as ($x_i(r)$, $y_j(r)$, $Z(r)$, r). For example, suppose $\mathcal{X}$ = {x_1, x_2, x_3) and $\mathcal{Y}$ = {y_1, y_2), and dyadic sample of 4 co-occurrences, $\mathcal{S}$ = {(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 x_k, let $\mathcal{S}_{x_k}$ be the $\mathcal{X}$-partitioned subset of $\mathcal{S}$ which contains co-occurrences whose $\mathcal{X}$-objects are fixed at x_k (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 1). Note, $\mathcal{S}_{x_k}$ can be empty. The size of $\mathcal{S}_{x_k}$ is $|\mathcal{S}_{x_k}|$.

$$\mathcal{S}_{x_k} = \{(x_i, y_j, z, r): x_i = x_k\} \tag{2.3}$$

Dyadic data $\mathcal{S}$ is partitioned into $|\mathcal{X}|$ subsets $\mathcal{S}_{x_k}$.

$$\mathcal{S} = \bigcup_{k=1}^{|\mathcal{X}|} \mathcal{S}_{x_k}$$
$$\forall i \neq j, \mathcal{S}_{x_i} \cap \mathcal{S}_{x_j} = \emptyset$$

Given fixed y_l, let $\mathcal{S}_{y_l}$ be the $\mathcal{Y}$-partitioned subset of $\mathcal{S}$ which contains co-occurrences whose $\mathcal{Y}$-objects are fixed at y_l. Note, $\mathcal{S}_{y_l}$ can be empty. The size of $\mathcal{S}_{y_l}$ is $|\mathcal{S}_{y_l}|$.

$$\mathcal{S}_{y_l} = \{(x_i, y_j, z, r): y_j = y_l\} \tag{2.4}$$

Dyadic data $\mathcal{S}$ is partitioned into $|\mathcal{Y}|$ subsets $\mathcal{S}_{y_l}$.

$$\mathcal{S} = \bigcup_{l=1}^{|\mathcal{Y}|} \mathcal{S}_{y_l}$$
$$\forall i \neq j, \mathcal{S}_{y_i} \cap \mathcal{S}_{y_j} = \emptyset$$

Given fixed x_k and fixed y_l, let $\mathcal{S}_{x_k y_l}$ be the subset of the $\mathcal{S}$ which contains co-occurrences whose $\mathcal{X}$-objects and $\mathcal{Y}$-objects are fixed at x_k and y_l. Note, $\mathcal{S}_{x_k y_l}$ can be empty. The size of $\mathcal{S}_{x_k y_l}$ is $\left|\mathcal{S}_{x_k y_l}\right|$.

$$\mathcal{S}_{x_k y_l} = \{(x_i, y_j, z, r): x_i = x_k, y_j = y_l\} \tag{2.5}$$

Let $n(x_i)$ and $n(y_j)$ denote the number of x_i and the number of y_j, respectively.

$$\begin{aligned} n(x_i) &= \left|\mathcal{S}_{x_i}\right| \\ n(y_j) &= \left|\mathcal{S}_{y_j}\right| \end{aligned} \tag{2.6}$$

Let $n(x_i, y_j)$ denote the number of co-occurrences (x_i, y_j).

$$n(x_i, y_j) = \left|\mathcal{S}_{x_i y_j}\right| \tag{2.7}$$

Let $n(x_i|y_j)$ and $n(y_j|x_i)$ denote the frequency of x_i given y_j and the frequency of y_j given x_i, respectively.

$$\begin{aligned} n(x_i|y_j) &= \frac{n(x_i, y_j)}{n(y_j)} \\ n(y_j|x_i) &= \frac{n(x_i, y_j)}{n(x_i)} \end{aligned} \tag{2.8}$$

For example, suppose $\mathcal{X} = \{x_1, x_2, x_3\}$ and $\mathcal{Y} = \{y_1, y_2\}$, and dyadic data of 4 co-occurrences, $\mathcal{S} = \{(x_1, y_1, 1), (x_1, y_1, 2), (x_1, y_2, 3), (x_1, y_1, 4)\}$, we have $\mathcal{S}_{x_1} = \{(x_1, y_1, 1), (x_1, y_1, 2), (x_1, y_2, 3), (x_1, y_1, 4)\}$, $\mathcal{S}_{x_2} = \mathcal{S}_{x_3} = \emptyset$, $\mathcal{S}_{y_1} = \{(x_1, y_1, 1), (x_1, y_1, 2), (x_1, y_1, 4)\}$, $\mathcal{S}_{y_2} = \{(x_1, y_2, 3)\}$, $\mathcal{S}_{x_1 y_1} = = \{(x_1, y_1, 1), (x_1, y_1, 2), (x_1, y_1, 4)\}$, $\mathcal{S}_{x_1 y_2} = \{(x_1, y_2, 3)\}$, $\mathcal{S}_{x_2 y_1} = \mathcal{S}_{x_2 y_2} = \mathcal{S}_{x_3 y_1} = \mathcal{S}_{x_3 y_2} = \emptyset$, $n(x_1) = 1$, $n(x_2) = n(x_3) = 0$, $n(y_1) = 3$, $n(y_2) = 1$, $n(x_1, y_1) = 3$, $n(x_1, y_2) = 1$, $n(x_2, y_1) = n(x_2, y_2) = n(x_3, y_1) = n(x_3, y_2) = 0$, $n(x_1 \mid y_1) = 1$, $n(x_1 \mid y_2) = 1$, $n(x_2 \mid y_1) = n(x_2 \mid y_2) = n(x_3 \mid y_1) = n(x_3 \mid y_2) = 0$, $n(y_1 \mid x_1) = 3/4$, $n(y_2 \mid x_1) = 1/4$.

Suppose each co-occurrence (x_i, y_j) belongs to a latent variable C and C has K values c_k (s). These values c_k (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 $c_k = k$ where $k = 1, 2, \ldots, K$. The random variable C has discrete distribution such that every value has an associated probability α_k. Of course, there are K probabilities α_k (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). This section only explains these models when they

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 α_k (s) are independent from both x_i and y_j. SMM is defined as follows (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 2):

$$P(x_i, y_j|\Theta) = \sum_{k=1}^{K} \alpha_k P(x_i, y_j|k) = \sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k} \tag{2.9}$$

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

$$\alpha_k = P(k)$$

The $p_{i|k}$ is the probability of x_i given aspect *k*.

$$p_{i|k} = P(x_i|k)$$

The $q_{j|k}$ is the probability of y_j given aspect *k*.

$$q_{j|k} = P(y_j|k)$$

This implies that x_i and y_j are mutually independent in SMM.

$$P(x_i, y_j|k) = P(x_i|k)P(y_j|k)$$

The joint probability of x_i, y_j, and *k* is:

$$P(x_i, y_j, k) = P(k)P(x_i, y_j|k) = \alpha_k P(x_i|k)P(y_j|k) = \alpha_k p_{i|k} q_{j|k}$$

The parameter of SMM is $\Theta = (\alpha_k, p_{i|k}, q_{j|k})^T$ in which there are $K(|\mathcal{X}| + |\mathcal{Y}| + 1)$ partial parameters α_k, $p_{i|k}$, and $q_{j|k}$. Note,

$$\sum_{k=1}^{K} \alpha_k = 1, \sum_{i=1}^{|\mathcal{X}|} p_{i|k} = 1, \sum_{j=1}^{|\mathcal{Y}|} q_{j|k} = 1$$

By applying GEM, given dyadic sample $\mathcal{S}$, at the t^{th} iteration of GEM, given current parameter $\Theta^{(t)} = (\alpha_k^{(t)}, p_{i|k}^{(t)}, q_{j|k}^{(t)})^T$, the conditional expectation $Q(\Theta|\Theta^{(t)})$ is:

$$Q(\Theta|\Theta^{(t)}) = \sum_{r=1}^{|\mathcal{S}|}\sum_{k=1}^{K} P(k|x_i(r), y_j(r), \Theta^{(t)})\log(\alpha_k p_{i|k} q_{j|k}) = \sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)\sum_{k=1}^{K} P(k|x_i, y_j, \Theta^{(t)})\left(\log(\alpha_k) + \log(p_{i|k}) + \log(q_{j|k})\right) \quad (2.10)$$

Where,

$$P(k|x_i, y_j, \Theta^{(t)}) = \frac{\alpha_k^{(t)} p_{i|k}^{(t)} q_{j|k}^{(t)}}{\sum_{l=1}^{K} \alpha_l^{(t)} p_{i|l}^{(t)} q_{j|l}^{(t)}} \quad (2.11)$$

Note, $n(x_i, y_j)$ is the number of co-occurrences (x_i, y_j) in $\mathcal{S}$, which is specified by equation 2.7. Please refer to equation 1.4 to comprehend equation 2.11. Because there are three constraints

$$\sum_{k=1}^{K} \alpha_k = 1, \sum_{i=1}^{|\mathcal{X}|} p_{i|k} = 1, \sum_{j=1}^{|\mathcal{Y}|} q_{j|k} = 1$$

We use Lagrange duality method to maximize $Q(\Theta|\Theta^{(t)})$. The Lagrange function $la(\Theta, \lambda \mid \Theta^{(t)})$ is sum of $Q(\Theta|\Theta^{(t)})$ and these constraints, as follows:

$$la(\Theta, \lambda|\Theta^{(t)}) = Q(\Theta|\Theta^{(t)}) + \lambda_1\left(1 - \sum_{k=1}^{K}\alpha_k\right) + \lambda_2\left(1 - \sum_{i=1}^{|\mathcal{X}|} p_{i|k}\right) + \lambda_3\left(1 - \sum_{j=1}^{|\mathcal{Y}|} q_{j|k}\right)$$
$$= \sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)\sum_{k=1}^{K} P(k|x_i, y_j, \Theta^{(t)})\left(\log(\alpha_k) + \log(p_{i|k}) + \log(q_{j|k})\right)$$
$$+ \lambda_1\left(1 - \sum_{k=1}^{K}\alpha_k\right) + \lambda_2\left(1 - \sum_{i=1}^{|\mathcal{X}|} p_{i|k}\right) + \lambda_3\left(1 - \sum_{j=1}^{|\mathcal{Y}|} q_{j|k}\right)$$

Note, $\lambda = (\lambda_1, \lambda_2, \lambda_3)^T$ where $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_3 \geq 0$ are called Lagrange multipliers. Of course, $la(\Theta, \lambda \mid \Theta^{(t)})$ is function of Θ and λ. The next parameters $\Theta^{(t+1)}$ that maximizes $Q(\Theta|\Theta^{(t)})$ 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 α_k is:

$$\frac{\partial la(\Theta, \lambda|\Theta^{(t)})}{\partial \alpha_k} = \sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)\frac{1}{\alpha_k} P(k|x_i, y_j, \Theta^{(t)}) - \lambda_1$$

Setting this partial derivative to be zero, we obtain:

$$\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \alpha_k \lambda_1 = 0$$

Summing the equation above over K aspects $\{1, 2,\ldots, K\}$, we have:

$$\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) \sum_{k=1}^{K} P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \lambda_1 \sum_{k=1}^{K} \alpha_k = 0$$

$$\Leftrightarrow \sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) - \lambda_1 = 0 \Leftrightarrow \lambda_1 = \sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)$$

This means the next parameters $\alpha_k^{(t+1)}$ is:

$$\alpha_k^{(t+1)} = \frac{\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)}{\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)} \tag{2.12}$$

The first-order partial derivative of Lagrange function regarding $p_{i|k}$ is:

$$\frac{\partial la\left(\Theta, \lambda \middle| \Theta^{(t)}\right)}{\partial p_{i|k}} = \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) \frac{1}{p_{i|k}} P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \lambda_2$$

Setting this partial derivative to be zero, we obtain:

$$\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - p_{i|k} \lambda_2 = 0$$

Summing the equation above over $\mathcal{X}$, we have:

$$\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \lambda_2 \sum_{i=1}^{|\mathcal{X}|} p_{i|k} = 0$$

$$\Leftrightarrow \lambda_2 = \sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)$$

This means the next parameters $p_{i|k}^{(t+1)}$ is:

$$p_{i|k}^{(t+1)} = \frac{\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)}{\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)} \tag{2.13}$$

Similarly, the next parameters $q_{j|k}^{(t+1)}$ is:

$$q_{j|k}^{(t+1)} = \frac{\sum_{i=1}^{|\mathcal{X}|} n(x_i, y_j) P(k|x_i, y_j, \Theta^{(t)})}{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P(k|x_i, y_j, \Theta^{(t)})} \quad (2.14)$$

The two steps of GEM algorithm for SMM at some t^{th} iteration are shown in Table 2.1.

Table 2.1: E-step and M-step of GEM algorithm for SMM.

E-step:

The conditional probability $P(k \mid x_i, y_j, \Theta^{(t)})$ is calculated based on current parameter $\Theta^{(t)} = (\alpha_k^{(t)}, p_{i|k}^{(t)}, q_{j|k}^{(t)})^T$, according to equation 2.11.

$$P(k|x_i, y_j, \Theta^{(t)}) = \frac{\alpha_k^{(t)} p_{i|k}^{(t)} q_{j|k}^{(t)}}{\sum_{l=1}^{K} \alpha_l^{(t)} p_{i|l}^{(t)} q_{j|l}^{(t)}}$$

M-step:

The next parameter $\Theta^{(t+1)} = (\alpha_k^{(t+1)}, p_{i|k}^{(t+1)}, q_{j|k}^{(t+1)})^T$, which is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with subject to Θ, is calculated by equation 2.12, equation 2.13, and equation 2.14.

$$\alpha_k^{(t+1)} = \frac{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P(k|x_i, y_j, \Theta^{(t)})}{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)}$$

$$p_{i|k}^{(t+1)} = \frac{\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P(k|x_i, y_j, \Theta^{(t)})}{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P(k|x_i, y_j, \Theta^{(t)})}$$

$$q_{j|k}^{(t+1)} = \frac{\sum_{i=1}^{|\mathcal{X}|} n(x_i, y_j) P(k|x_i, y_j, \Theta^{(t)})}{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P(k|x_i, y_j, \Theta^{(t)})}$$

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

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

$$P(x_i, y_j|\Theta) = p_i q_{j|i} = p_i \sum_{k=1}^{K} \alpha_{k|i} q_{j|k} \quad (2.15)$$

The $\alpha_{k|i}$ is the probability of aspect k given x_i.

$$\alpha_{k|i} = P(k|x_i)$$

Where p_i is the probability of x_i.

$$p_i = P(x_i)$$

The $q_{j|k}$ is the conditional probability of y_j given aspect k. Suppose y_j is independent from x_i given k, we have:

$$q_{j|k} = P(y_j|x_i, k) = P(y_j|k)$$

Note, $q_{j|i}$ is the conditional probability of y_j given x_i, which is defined as follows:

$$q_{j|i} = P(y_j|x_i) = \sum_{k=1}^{K} \alpha_{k|i} q_{j|k}$$

The joint probability of x_i, y_j, and k is:

$$P(x_i, y_j, k) = P(x_i)P(y_j, k|x_i) = P(x_i)P(k|x_i)P(y_j|x_i, k) = p_i \alpha_{k|i} P(y_j|k) = p_i \alpha_{k|i} q_{j|k}$$

The parameter of AMM is $\Theta = (\alpha_{k|i}, p_i, q_{j|k})^T$ in which there are $K(|\mathcal{X}| + |\mathcal{Y}|) + |\mathcal{X}|$ partial parameters $\alpha_{k|i}$, p_i, and $q_{j|k}$. Note,

$$\sum_{k=1}^{K} \alpha_{k|i} = 1, \sum_{i=1}^{|\mathcal{X}|} p_i = 1, \sum_{j=1}^{|\mathcal{Y}|} q_{j|k} = 1$$

By applying GEM, given dyadic sample $\mathcal{S}$, at the t^{th} iteration of GEM, given current parameter $\Theta^{(t)} = (\alpha_k^{(t)}, p_{i|k}^{(t)}, q_{j|k}^{(t)})^T$, the conditional expectation $Q(\Theta|\Theta^{(t)})$ is:

$$\begin{aligned} Q(\Theta|\Theta^{(t)}) &= \sum_{r=1}^{|\mathcal{S}|} \sum_{k=1}^{K} P(k|x_i(r), y_j(r), \Theta^{(t)}) \log(\alpha_{k|i} p_i q_{j|k}) \\ &= \sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) \sum_{k=1}^{K} P(k|x_i, y_j, \Theta^{(t)}) \left(\log(\alpha_{k|i}) + \log(p_i) \right. \\ &\quad \left. + \log(q_{j|k})\right) \end{aligned} \tag{2.16}$$

Where,

$$P(k|x_i, y_j, \Theta^{(t)}) = \frac{\alpha_{k|i}^{(t)} p_i^{(t)} q_{j|k}^{(t)}}{\sum_{l=1}^{K} \alpha_{l|i}^{(t)} p_i^{(t)} q_{j|l}^{(t)}} \tag{2.17}$$

Please refer to equation 1.4 to comprehend equation 2.17. Because there are three constraints

$$\sum_{k=1}^{K} \alpha_{k|i} = 1, \sum_{i=1}^{|\mathcal{X}|} p_i = 1, \sum_{j=1}^{|\mathcal{Y}|} q_{j|k} = 1$$

We use Lagrange duality method to maximize $Q(\Theta|\Theta^{(t)})$. The Lagrange function $la(\Theta, \lambda \mid \Theta^{(t)})$ is sum of $Q(\Theta|\Theta^{(t)})$ and these constraints, as follows:

$$\begin{aligned} la\left(\Theta, \lambda \middle| \Theta^{(t)}\right) &= Q\left(\Theta \middle| \Theta^{(t)}\right) + \lambda_1 \left(1 - \sum_{k=1}^{K} \alpha_{k|i}\right) + \lambda_2 \left(1 - \sum_{i=1}^{|\mathcal{X}|} p_i\right) + \lambda_3 \left(1 - \sum_{j=1}^{|\mathcal{Y}|} q_{j|k}\right) \\ &= \sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) \sum_{k=1}^{K} P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) \left(\log(\alpha_{k|i}) + \log(p_i) + \log(q_{j|k})\right) \\ &+ \lambda_1 \left(1 - \sum_{k=1}^{K} \alpha_{k|i}\right) + \lambda_2 \left(1 - \sum_{i=1}^{|\mathcal{X}|} p_i\right) + \lambda_3 \left(1 - \sum_{j=1}^{|\mathcal{Y}|} q_{j|k}\right) \end{aligned}$$

Note, $\lambda = (\lambda_1, \lambda_2, \lambda_3)^T$ where $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_3 \geq 0$ are called Lagrange multipliers. Of course, $la(\Theta, \lambda \mid \Theta^{(t)})$ is function of Θ and λ. The next parameters $\Theta^{(t+1)}$ that maximizes $Q(\Theta|\Theta^{(t)})$ 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 $\alpha_{k|i}$ is:

$$\frac{\partial la\left(\Theta, \lambda \middle| \Theta^{(t)}\right)}{\partial \alpha_{k|i}} = \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) \frac{1}{\alpha_{k|i}} P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \lambda_1$$

Setting this partial derivative to be zero, we obtain:

$$\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \alpha_{k|i} \lambda_1 = 0$$

Summing the equation above over K aspects $\{1, 2, \ldots, K\}$, we have:

$$\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) \sum_{k=1}^{K} P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \lambda_1 \sum_{k=1}^{K} \alpha_k = 0$$

$$\Leftrightarrow \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) - \lambda_1 = 0 \Leftrightarrow \lambda_1 = \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)$$

This means the next parameters $\alpha_{k|i}^{(t+1)}$ is:

$$\alpha_{k|i}^{(t+1)} = \frac{\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)}{\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)} \tag{2.18}$$

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

$$\frac{\partial la\left(\Theta, \lambda \middle| \Theta^{(t)}\right)}{\partial p_i} = \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) \frac{1}{p_i} - \lambda_2$$

Setting this partial derivative to be zero, we obtain:

$$\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) - p_i \lambda_2 = 0$$

Summing the equation above over $\mathcal{X}$, we have:

$$\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) - \lambda_2 \sum_{i=1}^{|\mathcal{X}|} p_i = 0$$

$$\Leftrightarrow \lambda_2 = \sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)$$

This means the next parameters $p_i^{(t+1)}$ is:

$$p_i^{(t+1)} = \frac{\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)}{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)} \tag{2.19}$$

The first-order partial derivative of Lagrange function regarding $q_{j|k}$ is:

$$\frac{\partial la\left(\Theta, \lambda \middle| \Theta^{(t)}\right)}{\partial q_{j|k}} = \sum_{i=1}^{|\mathcal{X}|} n(x_i, y_j) \frac{1}{q_{j|k}} P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \lambda_3$$

Setting this partial derivative to be zero, we obtain:

$$\sum_{i=1}^{|\mathcal{X}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - q_{j|k} \lambda_3 = 0$$

Summing the equation above over $\mathcal{Y}$, we have:

$$\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \lambda_3 \sum_{j=1}^{|\mathcal{Y}|} q_{j|k}$$

$$\Leftrightarrow \sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) - \lambda_3 \Leftrightarrow \lambda_3 = \sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)$$

This means the next parameters $q_{j|k}^{(t+1)}$ is:

$$q_{j|k}^{(t+1)} = \frac{\sum_{i=1}^{|\mathcal{X}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)}{\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)} \tag{2.20}$$

The two steps of GEM algorithm for AMM at some t^{th} iteration are shown in Table 2.2.

Table 2.2: E-step and M-step of GEM algorithm for AMM.

E-step:

The conditional probability $P(k \mid x_i, y_j, \Theta^{(t)})$ is calculated based on current parameter $\Theta^{(t)} = (\alpha_{k|i}^{(t)}, p_i^{(t)}, q_{j|k}^{(t)})^T$, according to equation 2.17.

$$P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) = \frac{\alpha_{k|i}^{(t)} p_i^{(t)} q_{j|k}^{(t)}}{\sum_{l=1}^{K} \alpha_{l|i}^{(t)} p_i^{(t)} q_{j|l}^{(t)}}$$

M-step:

The next parameter $\Theta^{(t+1)} = (\alpha_{k|i}^{(t+1)}, p_i^{(t+1)}, q_{j|k}^{(t+1)})^T$, which is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with subject to Θ, is calculated by equation 2.18, equation 2.19, and equation 2.20.

$$\alpha_{k|i}^{(t+1)} = \frac{\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)}{\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)}$$

$$p_i^{(t+1)} = \frac{\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)}{\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)}$$

$$q_{j|k}^{(t+1)} = \frac{\sum_{i=1}^{|\mathcal{X}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)}{\sum_{i=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)}$$

GEM algorithm converges at some t^{th} iteration. At that time, $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ is the AMM itself. When AMM is applied into soft clustering, dyadic data is clustered vertically (horizontally) and each $\alpha_{k|i}$ is

coverage ratio of cluster *k* (aspect *k*) according to x_i. 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 $\mathcal{X}$-aspect set {1, 2,…, $K_\mathcal{X}$} and $\mathcal{Y}$-aspect set {1, 2,…, $K_\mathcal{Y}$}. In other words, the aspect space is still symmetric but is checked (stripped) according to two directions $\mathcal{X}$ and $\mathcal{Y}$.

$$\begin{aligned} &\{1,2,\dots,K\} \sim \{1,2,\dots,K_\mathcal{X}\} \times \{1,2,\dots,K_\mathcal{Y}\} \\ &K = K_\mathcal{X} K_\mathcal{Y} \end{aligned} \tag{2.21}$$

For every *k* belongs to {1, 2,…, *K*}, there always exists a respective pair: $k_\mathcal{X} \in \{1,2,\dots,K_\mathcal{X}\}$ and $k_\mathcal{Y} \in \{1,2,\dots,K_\mathcal{Y}\}$. However, for each $k_\mathcal{X}$ or each $k_\mathcal{Y}$, there are many respective *k*.

$$\begin{aligned} &k \sim \{k_\mathcal{X}, k_\mathcal{Y}\} \\ &k_\mathcal{X} \sim \text{many } k \\ &k_\mathcal{Y} \sim \text{many } k \end{aligned} \tag{2.22}$$

The sign "∼" denotes correspondence. PMM is defined as follows (Hofmann & Puzicha, Statistical Models for Co-occurrence Data, 1998, p. 4):

$$P(x_i, y_j|\Theta) = \sum_{k=1}^{K} \alpha_k p_{i|k_\mathcal{X}} q_{j|k_\mathcal{Y}} \tag{2.23}$$

As usual, α_k is the probability of aspect c_k but $p_{i|k_\mathcal{X}}$ is the probability of x_i given $k_\mathcal{X}$ of *k* and $q_{j|k_\mathcal{Y}}$ is the probability of y_j given $k_\mathcal{Y}$ of *k*.

$$\begin{aligned} p_{i|k_\mathcal{X}} &= P(x_i|k_\mathcal{X}) \\ q_{j|k_\mathcal{Y}} &= P(y_j|k_\mathcal{Y}) \end{aligned}$$

The joint probability of x_i, y_j, and *k* is:

$$P(x_i, y_j, k) = P(k)P(x_i, y_j|k) = \alpha_k P(x_i|k)P(y_j|k) = \alpha_k P(x_i|k_\mathcal{X})P(y_j|k_\mathcal{Y}) = \alpha_k p_{i|k_\mathcal{X}} q_{j|k_\mathcal{Y}}$$

The parameter of PMM is $\Theta = (\alpha_k, p_{i|k_\mathcal{X}}, q_{j|k_\mathcal{Y}})^T$ in which there are $K + K_\mathcal{X}|\mathcal{X}| + K_\mathcal{Y}|\mathcal{Y}|$ partial parameters α_k, $p_{i|k_\mathcal{X}}$, and $q_{j|k_\mathcal{Y}}$. Note,

$$\sum_{k=1}^{K} \alpha_k = 1, \sum_{i=1}^{|\mathcal{X}|} p_{i|k_\mathcal{X}} = 1, \sum_{j=1}^{|\mathcal{Y}|} q_{j|k_\mathcal{Y}} = 1$$

Learning PMM is like learning SMM and so it is not necessary to duplicate the expansion of $Q(\Theta|\Theta^{(t)})$. The two steps of GEM algorithm for PMM at some t^{th} iteration are shown in Table 2.3.

Table 2.3: E-step and M-step of GEM algorithm for PMM.

E-step:

The conditional probabilities $P(k \mid x_i, y_j, \Theta^{(t)})$, $P(k_{\mathcal{X}} \mid x_i, y_j, \Theta^{(t)})$, and $P(k_{\mathcal{Y}} \mid x_i, y_j, \Theta^{(t)})$ are calculated based on current parameter $\Theta^{(t)} = \left(\alpha_k^{(t)}, p_{i|k_{\mathcal{X}}}^{(t)}, q_{j|k_{\mathcal{Y}}}^{(t)}\right)^T$, according to equation 2.24, equation 2.25, and equation 2.26.

$$P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) = \frac{\alpha_k^{(t)} p_{i|k_{\mathcal{X}}}^{(t)} q_{j|k_{\mathcal{Y}}}^{(t)}}{\sum_{l=1}^{K} \alpha_l^{(t)} p_{i|l_{\mathcal{X}}}^{(t)} q_{j|l_{\mathcal{Y}}}^{(t)}} \tag{2.24}$$

$$P\left(k_{\mathcal{X}} \middle| x_i, y_j, \Theta^{(t)}\right) = \sum_{k:k_{\mathcal{X}} \sim k} P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) \tag{2.25}$$

$$P\left(k_{\mathcal{Y}} \middle| x_i, y_j, \Theta^{(t)}\right) = \sum_{k:k_{\mathcal{Y}} \sim k} P\left(k \middle| x_i, y_j, \Theta^{(t)}\right) \tag{2.26}$$

Please refer to equation 1.4 to comprehend equation 2.24.

M-step:

The next parameter $\Theta^{(t+1)} = \left(\alpha_k^{(t+1)}, p_{i|k_{\mathcal{X}}}^{(t+1)}, q_{j|k_{\mathcal{Y}}}^{(t+1)}\right)^T$, which is the maximizer of $Q(\Theta \mid \Theta^{(t)})$ with subject to Θ, is calculated by equation 2.27, equation 2.28, and equation 2.29.

$$\alpha_k^{(t+1)} = \frac{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k \middle| x_i, y_j, \Theta^{(t)}\right)}{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j)} \tag{2.27}$$

$$p_{i|k_{\mathcal{X}}}^{(t+1)} = \frac{\sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k_{\mathcal{X}} \middle| x_i, y_j, \Theta^{(t)}\right)}{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k_{\mathcal{X}} \middle| x_i, y_j, \Theta^{(t)}\right)} \tag{2.28}$$

$$q_{j|k_{\mathcal{Y}}}^{(t+1)} = \frac{\sum_{i=1}^{|\mathcal{X}|} n(x_i, y_j) P\left(k_{\mathcal{Y}} \middle| x_i, y_j, \Theta^{(t)}\right)}{\sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} n(x_i, y_j) P\left(k_{\mathcal{Y}} \middle| x_i, y_j, \Theta^{(t)}\right)} \tag{2.29}$$

GEM algorithm converges at some t^{th} iteration. At that time, $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ is the PMM itself. When PMM is applied into soft clustering, dyadic data is clustered in checked (stripped) and each α_k is

coverage ratio of cluster *k* (aspect *k*) but such cluster *k* corresponds to a pair of cluster $k_{\mathcal{X}}$ and cluster $k_{\mathcal{Y}}$. Soft clustering with PMM is also called two-side clustering.

3. Predicting unaccomplished co-occurrent values

This section is the main subject of this research in which some extensions of dyadic mixture models are used to predict unaccomplished values in valued dyadic data. When $\mathcal{S}$ is valued dyadic data in which every co-occurrence (x_i, y_j) is associated with value *z* from random variable *Z* then, SMM is reformed as follows:

$$f(x_i, y_j, Z|\Theta) = \sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k} f_k(Z|\varphi_k) \tag{3.1}$$

AMM is reformed as follows:

$$f(x_i, y_j, Z|\Theta) = p_i \sum_{k=1}^{K} \alpha_{k|i} q_{j|k} f_k(Z|\varphi_k) \tag{3.2}$$

PMM is reformed as follows:

$$f(x_i, y_j, Z|\Theta) = \sum_{k=1}^{K} \alpha_k p_{i|k_{\mathcal{X}}} q_{j|k_{\mathcal{Y}}} f_k(Z|\varphi_k) \tag{3.3}$$

Where $f_k(Z|\varphi_k)$ is the k^{th} PDF of *Z* corresponding to the aspect *k*, in which φ_k is parameter of $f_k(Z|\varphi_k)$. Of course, the parameter Θ now must include all φ_k. It is possible to consider that

$$f_k(Z|\varphi_k) = f(Z|k, \varphi_k)$$

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

$$f(Z|x_i, k, \varphi_k) = f(Z|k, \varphi_k) = f_k(Z|\varphi_k)$$

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

$$P(x_i, y_j|k, Z) = P(x_i, y_j|k)$$

$$P(y_j|x_i, Z, k) = P(y_j|x_i, k)$$

For example, within SMM, the joint PDF of x_i, y_j, *Z*, and *k* is:

$$f(x_i, y_j, Z, k) = P(k)P(x_i, y_j, Z|k) = \alpha_k P(x_i, y_j|k, Z) f(Z|k, \varphi_k) = \alpha_k P(x_i, y_j|k) f_k(Z|\varphi_k)$$
$$= \alpha_k P(x_i|k) P(y_j|k) f_k(Z|\varphi_k) = \alpha_k p_{i|k} q_{j|k} f_k(Z|\varphi_k)$$

Within AMM, the joint PDF of x_i, y_j, Z, and k is:

$$f(x_i, y_j, Z, k) = P(x_i)P(y_j, Z, k|x_i) = p_i P(k|x_i)P(y_j, Z|x_i, k) = p_i \alpha_{k|i} P(y_j|x_i, Z, k) f(Z|x_i, k, \varphi_k)$$
$$= p_i \alpha_{k|i} P(y_j|x_i, k) f(Z|k, \varphi_k) = p_i \alpha_{k|i} P(y_j|k) f_k(Z|\varphi_k) = p_i \alpha_{k|i} q_{j|k} f_k(Z|\varphi_k)$$

Within PMM, the joint PDF of x_i, y_j, Z, and k is:

$$f(x_i, y_j, Z, k) = P(k)P(x_i, y_j, Z|k) = \alpha_k P(x_i, y_j|Z, k) f(Z|k, \varphi_k) = \alpha_k P(x_i, y_j|k) f_k(Z|\varphi_k)$$
$$= \alpha_k P(x_i|k_X) P(y_j|k_Y) f_k(Z|\varphi_k) = \alpha_k p_{i|k_X} q_{j|k_Y} f_k(Z|\varphi_k) \blacksquare$$

Here it is only necessary to estimate φ_k because how to estimate other partial parameters was mentioned in section 2. By reforming the conditional expectation $Q(\Theta|\Theta^{(t)})$, it is easy to find out that the next parameter $\varphi_k^{(t+1)}$ is solution of following equation:

$$\sum_{r=1}^{|\mathcal{S}|} P\left(k|x_i(r), y_j(r), \Theta^{(t)}\right) \frac{\mathrm{d}\log\left(f_k(Z(r)|\varphi_k)\right)}{\mathrm{d}\varphi_k} \tag{3.4}$$

Where $P(k \mid x_i(r), y_j(r), \Theta^{(t)})$ is specified by equation 2.11, equation 2.17, and equation 2.24 for SMM, AMM, and PMM, respectively. Especially, if $f_k(Z|\varphi_k)$ distributed normally, the next parameter $\varphi_k^{(t+1)} = (\mu_k^{(t+1)}, \Sigma_k^{(t+1)})^T$ containing mean $\mu_k^{(t+1)}$ and covariance matrix $\Sigma_k^{(t+1)}$ is calculated as follows:

$$\mu_k^{(t+1)} = \frac{\sum_{r=1}^{|\mathcal{S}|} P\left(k|x_i(r), y_j(r), \Theta^{(t)}\right) Z(r)}{\sum_{i=1}^{|\mathcal{S}|} P\left(k|x_i(r), y_j(r), \Theta^{(t)}\right)}$$
$$\Sigma_k^{(t+1)} = \frac{\sum_{r=1}^{|\mathcal{S}|} P\left(k|x_i(r), y_j(r), \Theta^{(t)}\right) \left(\left(Z(r) - \mu_k^{(t+1)}\right)\left(Z(r) - \mu_k^{(t+1)}\right)^T\right)}{\sum_{r=1}^{|\mathcal{S}|} P\left(k|x_i(r), y_j(r), \Theta^{(t)}\right)} \tag{3.5}$$

Where $P(k \mid x_i(r), y_j(r), \Theta^{(t)})$ is specified by equation 2.11, equation 2.17, and equation 2.24 for SMM, AMM, and PMM, respectively. Please refer to (Nguyen, 2020, pp. 83-84) to comprehend equation 3.5.

In valued dyadic sample $\mathcal{S}$, many co-occurrences (x_i, y_j) are not existent and thus, it is required to predict or estimate Z value of inexistent co-occurrence (x_i, y_j). This Z value is called unaccomplished co-occurrent value or unaccomplished associative value. A so-called expected co-occurrent (EC) method is used to estimate Z. Firstly, it is necessary to define the conditional PDF of Z given x_i and y_j. According to Bayes' rule, we have:

$$f(Z|x_i, y_j, \Theta) = \frac{f(x_i, y_j, Z)}{\int_Z f(x_i, y_j, Z|\Theta)\mathrm{d}Z} = \frac{f(x_i, y_j, Z|\Theta)}{f(x_i, y_j|\Theta)} \tag{3.6}$$

Then, Z value of inexistent co-occurrence (x_i, y_j) is estimated by an estimate $\hat{Z}$ which is the expectation of Z given the conditional PDF $f(Z \mid x_i, y_j, \Theta)$, as follows:

$$\hat{Z} = E(Z|\Theta) = \int_Z Z f\left(Z\middle|x_i, y_j, \Theta\right)\mathrm{d}Z \tag{3.7}$$

In short, EC method is specified by equation 3.6 and equation 3.7. Now we expend the two equations for SMM, AMM, and PMM. The conditional PDF $f(Z \mid x_i, y_j, \Theta)$ of SMM is:

$$f\left(Z\middle|x_i, y_j, \Theta\right) = \frac{\sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k} f_k(Z|\varphi_k)}{\sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k}} \tag{3.8}$$

Following is the proof of equation 3.8.

$$f\left(Z\middle|x_i, y_j, \Theta\right) = \frac{\sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k} f_k(Z|\varphi_k)}{\int_Z \sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k} f_k(Z|\varphi_k)} = \frac{\sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k} f_k(Z|\varphi_k)}{\sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k} \int_Z f_k(Z|\varphi_k)}$$
$$= \frac{\sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k} f_k(Z|\varphi_k)}{\sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k}} \blacksquare$$

Similarly, the conditional PDF $f(Z \mid x_i, y_j, \Theta)$ of AMM is:

$$f\left(Z\middle|x_i, y_j, \Theta\right) = \frac{\sum_{k=1}^{K} \alpha_{k|i} q_{j|k} f_k(Z|\varphi_k)}{\sum_{k=1}^{K} \alpha_{k|i} q_{j|k}} \tag{3.9}$$

The conditional PDF $f(Z \mid x_i, y_j, \Theta)$ of PMM is:

$$f\left(Z\middle|x_i, y_j, \Theta\right) = \frac{\sum_{k=1}^{K} \alpha_k p_{i|k_x} q_{j|k_y} f_k(Z|\varphi_k)}{\sum_{k=1}^{K} \alpha_k p_{i|k_x} q_{j|k_y}} \tag{3.10}$$

Obviously, equation 3.8, equation 3.9, and equation 3.10 are extensions of equation 3.6.

The estimate $\hat{Z}$ for SMM is:

$$\hat{Z} = \frac{\sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k} E_k(Z|\varphi_k)}{\sum_{k=1}^{K} \alpha_k p_{i|k} q_{j|k}} \tag{3.11}$$

The estimate $\hat{Z}$ for AMM is:

$$\hat{Z} = \frac{\sum_{k=1}^{K} \alpha_{k|i} q_{j|k} E_k(Z|\varphi_k)}{\sum_{k=1}^{K} \alpha_{k|i} q_{j|k}} \tag{3.12}$$

The estimate $\hat{Z}$ for PMM is:

$$\hat{Z} = \frac{\sum_{k=1}^{K} \alpha_k p_{i|k_X} q_{j|k_Y} E_k(Z|\varphi_k)}{\sum_{k=1}^{K} \alpha_k p_{i|k_X} q_{j|k_Y}} \tag{3.13}$$

Where $E_k(Z|\varphi_k)$ is expectation of Z given the k^{th} PDF of Z:

$$E_k(Z|\varphi_k) = \int_Z Z f_k(Z|\varphi_k) \mathrm{d}Z \tag{3.14}$$

If $f_k(Z|\varphi_k)$ is multinormal PDF with mean μ_k and covariance matrix Σ_k then, we have $E_k(Z|\varphi_k) = \mu_k$. Note, equation 3.11, equation 3.12, and equation 3.13 are extensions of equation 3.7.

Hofmann's research (Hofmann, Latent Semantic Models for Collaborative Filtering, 2004) is different from EC method when Hofmann assumed that $f_k(Z|\varphi_k)$ is dependent on both k and x_i so that $f_k(Z|\varphi_k)$ is replaced by $f_{ik}(Z|\varphi_{ik})$.

$$f_{ik}(Z|\varphi_{ik}) = f(Z|x_i, k, \varphi_{ik}) = f\left(Z\middle|x_i, y_j, k, \varphi_{ik}\right)$$

Hofmann also assumed that (Hofmann & Puzieha, Latent Class Models for Collaborative Filtering, 1999, p. 690)

$$P\left(k\middle|x_i, y_j\right) = P\left(k\middle|y_j\right) = \frac{P(k)P\left(y_j\middle|k\right)}{\sum_{k=1}^{K} P(k)P\left(y_j\middle|k\right)} = \frac{\alpha_k q_{j|k}}{\sum_{k=1}^{K} \alpha_k q_{j|k}} \propto \alpha_k q_{j|k}$$

The sign "∝" indicates the proportion. Therefore, according to Hofmann, the conditional PDF $f(Z \mid x_i, y_j, \Theta)$ was defined as follows:

$$f\left(Z\middle|x_i, y_j, \Theta\right) = \sum_{k=1}^{K} P\left(k\middle|x_i, y_j\right) f\left(Z\middle|x_i, y_j, k, \varphi_{ik}\right) \propto \sum_{k=1}^{K} \alpha_k q_{j|k} f_{ik}(Z|\varphi_{ik}) \tag{3.15}$$

The estimate $\hat{Z}$ is still calculated by equation 3.7 except that $f(Z \mid x_i, y_j, \Theta)$ was defined by equation 3.15. As a result, equation 3.15 is the real mixture model of Hofmann in (Hofmann, Latent Semantic Models for Collaborative Filtering, 2004) and then Hofmann applied EM algorithm to learn parameters α_k, $q_{j|k}$, and φ_{ik}. Therefore, Hofmann's mixture model in (Hofmann, Latent Semantic Models for Collaborative Filtering, 2004) is not mixture models of co-occurrences (x_i, y_j) specified by equation 2.9 (SMM), equation 2.15 (AMM), and 2.23 (PMM). Hofmann's mixture model is appropriate to collaborative filtering.

4. Conclusions

Essentially, learning dyadic data with models such as SMM, AMM, and PMM is unsupervised learning and it is easy to apply these models into soft clustering. Predicting or estimating unaccomplished values is essential to make a weighted sum of centroids over all clusters. Currently, an unaccomplished value is estimated based on pre-knowledge of an existent pair of two objects ($\mathcal{X}$-object and $\mathcal{Y}$-object). As a

result, an estimate $\hat{Z}$ is fixed if the two objects are fixed. In future, we try to find out another method to take advantages of more than two existent objects with a set of values. Combination of dyadic mixture model and regression model is a candidate method but how to prove and explain it is still fuzzy problem.

Declaration of Interest Statement

An original version of this paper was published as the third chapter in the book "Some Applications of Expectation Maximization Algorithm" by Eliva Press (https://www.elivapress.com/en/book/book-5132639477) available at https://www.researchgate.net/publication/359661325_Some_Applications_of_Expectation_Maximization_Algorithm

References

Hofmann, T. (2004, January). Latent Semantic Models for Collaborative Filtering. *ACM Transactions on Information Systems (TOIS), 22*(1), 89-115. doi:10.1145/963770.963774

Hofmann, T., & Puzicha, J. (1998). *Statistical Models for Co-occurrence Data.* Massachusetts Institute of Technology, Artificial Intelligence Laboratory. MIT Publisher. Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/7253/AIM-1625.pdf?sequence=2

Hofmann, T., & Puzieha, J. (1999). Latent Class Models for Collaborative Filtering. In T. Dean (Ed.), *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI '99)* (pp. 688-693). San Francisco, CA, USA: Morgan Kaufmann. Retrieved from https://dl.acm.org/citation.cfm?id=687583

Hofmann, T., Puzicha, J., & Jordan, M. I. (1998). Learning from Dyadic Data. In M. J. Kearns, S. A. Solla, & D. A. Cohn (Ed.), *Advances in Neural Information Processing Systems 11 (NIPS 1998). 11*, pp. 466-472. Denver: MIT Press. Retrieved from https://papers.nips.cc/paper/1503-learning-from-dyadic-data

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